Appearance Modelling, Pathology Classification and Evidence Pinpointing for Medical Image Analysis

by

Qiang Zhang

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Computer Science

April 2017

Contents

Acknowledgments

First and foremost I would like to take the opportunity to express my deepest gratitude and respect to my supervisor, Dr. Abhir Bhalerao, who constantly offered fruitful guidance and strong support during all the time of my PhD study at the University of Warwick. I benefited greatly from his insightful advice, continuous encouragement and generous help. I am looking forward to our collaboration in the future. I am very grateful to Prof. Charles Hutchinson for the insightful advice and continuous support. I am impressed by his extensive knowledge in medicine as well as his kindness.

I wish to express my sincere thankfulness to Prof. Chang-Tsun Li and Dr. Victor Sanchez, for their guidance and valuable suggestions on my PhD progress.

I am thankful for the colleagues at the Warwick Medical School and University Hospitals Coventry and Warwickshire, Emma Helm, Caron Parsons and Edward Dickenson, for helping me with the clinical background knowledge. I would also like to thank the colleagues at the department, Xingjie Wei, Yu Guan, Yi Yao, Xufeng Lin, Ruizhe Li, Alaa Khadidos, Ning Jia, Roberto Leyva, Xin Guan, Bo Wang, Shan Lin, Ching-Chun Chang and Yijun Quan for their support and friendship.

Last but not least I would like to express my deepest gratitude to my parents for their love, understanding and encouragement throughout my life.

Declarations

I hereby declare that the work presented in this thesis entitled *Appearance Modelling, Pathology Classification and Evidence Pinpointing for Medical Image Analysis* is an original work and has not been submitted to any college, university or any other academic institution for the purpose of obtaining an academic degree.

Abbreviations

DAP Deformable Appearance Model

LK Lucas-Kanade

WAP Wavelet Appearance Pyramid

SDM Supervised Descent Method

RDA Regularised Dual Averaging

LSS Lumbar Spinal Stenosis

AAM Active Appearance Model

DPM Deformable Part Model

HOG Histogram of Oriented Gradient

CNN Convolutional Neural Network

LFP Local Feature Pyramid

DICOM Digital Imaging and Communications in Medicine

GUI Graphical User Interface

PCA Principal Component Analysis

CLM Cosntrained Local Model

KDE Kernel Density Estimate

CAD Computer-Aided Diagnosis

SVM Support Vector Machine

MIL Multiple-Instance Learning

ML Maximum Likelihood

SIFT Scale-Invariant Feature Transform

PtoBD Point to Boundary Distance

DSC Dice Similarity Coefficients

SD Standard Deviation

SNR Signal-to-Noise Ratio

MAE Mean Absolute Error

RMSE Root Mean Squared Error

L-SIFT Loglet-SIFT

DSP Domain Size Pooling

WFP Wavelet Feature Pyramid

LFPW Labeled Face Parts in the Wild

300-W 300 Faces In-the-Wild

Publications

Journal Articles

1. **Qiang Zhang**, Abhir Bhalerao, and Charles Hutchinson,: "Deformable Appearance Pyramids for Anatomy Representation, Landmark Detection and Pathology Classification", MICCAI Special issue, IJCARS, 2017.

2. **Qiang Zhang**, Abhir Bhalerao, Edward Dickenson, and Charles Hutchinson,: "Active appearance pyramids for object parametrisation and fitting." Medical Image Analysis (2016),
https://sites.google.com/site/activeappearancepyramids/.

Conference Papers

1. **Qiang Zhang**, Abhir Bhalerao, and Charles Hutchinson,: "Weakly supervised evidence pinpointing and description." Biennial International conference on Information Processing in Medical Imaging (IPMI 2017), North Carolina, USA, Jun 2017.

2. **Qiang Zhang**, Abhir Bhalerao, Caron Parsons, Emma Helm, and Charles Hutchinson,: "Wavelet Appearance Pyramids for Landmark Detection and Pathology Classification: Application to Lumbar Spinal Stenosis." Medical Image Computing and Computer Assisted Intervention (MICCAI 2016), Athens, Greece, Oct 2016,
https://sites.google.com/site/waveletappearancepyramids/.

3. **Qiang Zhang**, Abhir Bhalerao,: "Loglet SIFT for Part Description in Deformable Part Models: Application to Face Alignment." British Machine Vision Conference (BMVC 2016), York, Sep 2016,
   https://sites.google.com/site/logletsift/.

4. **Qiang Zhang**, Abhir Bhalerao, Emma Helm, and Charles Hutchinson,: "Active shape model unleashed with multi-scale local appearance." IEEE International Conference on Image Processing (ICIP 2015), Quebec City, Canada, Oct 2015.

5. Caron Parsons, Charles Hutchinson, Emma Helm, Alexander Clarke, Asfand Baig Mirza, **Qiang Zhang**, and Abhir Bhalerao,: "Development of an Automated Shape and Textural Software Model of the Paediatric Knee for Estimation of Skeletal Age." International Society for Magnetic Resonance in Medicine (ISMRM 2016), The 24th Annual Meeting and Exhibition, Singapore, May 2016.

Abstract

We propose several methods to address the tasks of appearance representation, variation modelling, landmark detection, pathology classification and evidence pinpointing in medical image analysis.

Object class representation is one of the key steps in various medical image understanding techniques. We propose a part-based parametric appearance model built on Gaussian pyramids we refer to as a Deformable Appearance Model (DAP). A DAP models the variability within a population with local translations of multi-scale parts and linear appearance variations of the assembly of the parts. The fitting process uses a two-step iterative strategy: local landmark searching followed by shape regularisation. We present a simultaneous local feature searching and appearance fitting algorithm based on the weighted Lucas-Kanade (LK) method. A shape regulariser is derived to calculate the maximum likelihood shape with respect to the prior and multiple landmark candidates from multi-scale parts, with a compact closed-form solution. We apply the DAP for the tasks of variation modelling and landmark detection.

To reduce the redundancy in the representation, we further propose to replace the Gaussian pyramids with wavelet pyramids in the DAPs. The new appearance model is referred to as a Wavelet Appearance Pyramid (WAP). Logarithmic wavelets are adopted to decompose the images into pyramidal complementary channels, each of which represents the image with simple textures at a given scale. The complementary property of the wavelets allows the reconstruction of the object

appearance from the image channels. The Supervised Descent Method (SDM) is adopted to model implicitly the prior knowledge and fit the model to new instances. We apply the WAPs for the tasks of landmark detection and pathology classification.

To learn on large scale datasets annotated with only class labels and no landmarks, we propose a weakly-supervised method utilising the theories of sparse learning and stochastic optimisation. We pay attention to identifying which specific regions and features of images contribute to a certain classification. In the medical imaging scenario, these can be the evidence regions where abnormalities are most likely to appear, and the discriminative features of these regions supporting the pathology classification. The learning is weakly-supervised requiring only the pathological labelling of the data by clinicians and no other prior knowledge. It can also be applied to learn the salient description of an anatomy discriminative from background, in order to localise the anatomy before a classification step. We formulate evidence pinpointing as a sparse descriptor learning problem. Because of the large computational complexity, the objective function is composed in a stochastic way and is optimised by the Regularised Dual Averaging (RDA) algorithm. We apply the evidence pinpointing method for the tasks of anatomy localisation and pathology classification.

We test our object representation and evidence pinpointing methods on the problem of Lumbar Spinal Stenosis (LSS). We validate the performance of DAPs and WAPs on around 200 studies consisting of routine axial and sagittal MRI scans. Intervertebral sagittal and parasagittal cross-sections are typically inspected for the diagnosis of LSS, we therefore build the appearance models on L3/4, L4/5 and L5/S1 axial cross-sections and parasagittal slices. For the task of landmark detection, experiments validate the performance of the DAPs as promising in terms of convergence range, robustness to local minima and segmentation precision compared with conventional shape and appearance models. A further improvement using WAPs is observed in landmark detection and pathology classification. We validate the evidence pinpointing method on three weakly annotated datasets on 600 axial images. Experiments show that compared with supervised methods trained with

labels and landmarks, our method gives favourable results trained on larger scale data with only class labels, which demonstrates the learning ability of our method under weak-supervision.

CHAPTER 1

Introduction

The advances in medical imaging over the last few decades have greatly improved the type of medical care that is available to patients [1]. Clinicians have the ability to non-invasively peer inside the human body to diagnose, treat, monitor changes, and plan and execute procedures more safely and effectively than before such medical imaging techniques existed. Imaging modalities [2] such as Magnetic Resonance (MR), Computed Tomography (CT), ultrasound, and Positron Emission Tomography (PET) reveal different information of the structure and function of internal anatomy. However, the imaging techniques in existence on their own are generally not able to directly provide the measures and information such as the precise location of anatomical landmarks, the pathology labels and the evidence of abnormalities. Advanced image processing and machine learning algorithms [3] are therefore in demand for understanding and extracting useful information from medical images.

A practical approach to understanding medical images may includes localising the anatomy of interest, classifying the pathology labels and revealing the evidence of pathologies. For a robust localisation of an anatomy, the prior knowledge of the shape and appearance is usually learnt through statistical appearance models. To classify the pathology condition, the features are extracted from the

appearances to train a classifier. In addition to predicting the labels, being able to pinpoint the evidence supporting the prediction can be extremely useful for more comprehensive evaluation. In this thesis we focus on three key tasks, namely, landmark detection, pathology classification and evidence pinpointing. We apply and validate the methods on the clinical problem of lumbar spinal stenosis.

1.1 Clinical Background on Lumbar Spinal Stenosis

Lumbar spinal stenosis (LSS) is the most frequent indication for spine surgery in older adults, and its prevalence is likely to increase. The diagnosis is most common among adults older than 65 years, a population projected to increase by 59% by 2025, to almost 65 million [4]. In clinical medicine lumbar spinal stenosis is defined as buttock or lower extremity pain, which may occur with or without back pain, associated with diminished space available for the neural and vascular elements in the lumbar spine [5]. This definition includes two aspects: morphological abnormalities and clinical symptoms. From a radiological perspective, emphasising the underlying structural anomaly, stenosis of the spinal canal with or without clinical manifestations is a more appropriate definition.

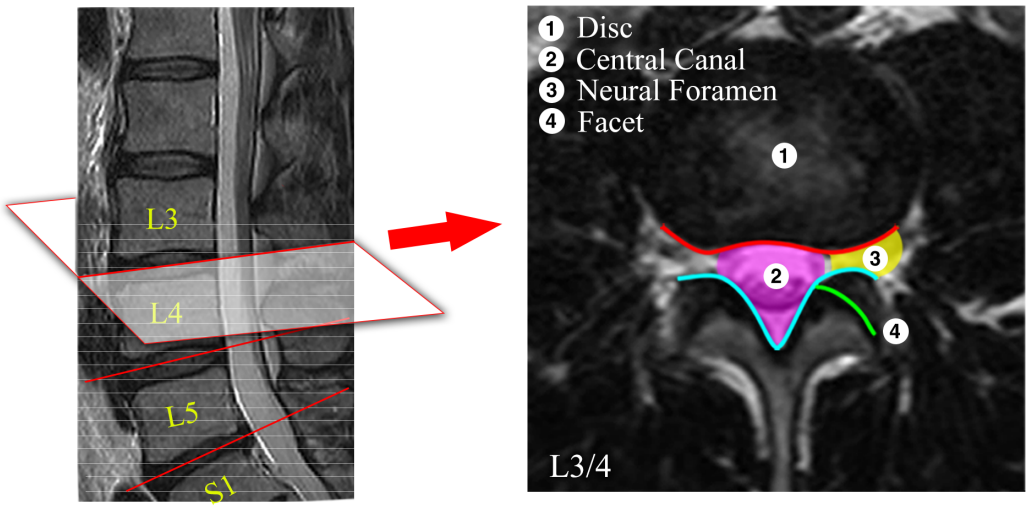
Proper research in patients with a particular illness requires a precise defination of the illness in order to formulate sensible and reliable inclusion criteria [6]. In a recent review [7] it is reported that researchers used quite different combinations of symptoms, clinical signs and radiological criteria to set up inclusion criteria for trials in patients with lumbar spinal stenosis. Imprecise definitions limit the interpretability of trial results as well as the efficiency in clinical practices. As LSS is a condition mostly caused by the morphological degeneration of the lumbar spine, radiological imaging is the preferential noninvasive test for tracing the lesions and evaluating the condition.

In MRI scans, the intervertebral disc-level axial views contain the richest

information of the important anatomical structures. The disc planes are localised from the sagittal scans (Fig. 1.1(a)), and the geometry is projected to the axial scans to extract the precise disc-level images. In an axial image (Fig. 1.1(b)), conditions of the posterior margins of the disc (red line), posterior spinal canal (cyan line) and the facet between the superior and inferior articular processes (green line) are typically evaluated for diagnosis and grading. Degeneration of these structures can constrict the spinal canal and the neural foramen causing central and foraminal stenosis. An example with central canal narrowing is given in Fig. 1.2(a), in which the spinal cord is suppressed by the disc with herniation. Fig. 1.2(b) shows a case with foraminal stenosis, in which the neural foramen is constricted by the thickening of the facet and the posterior margin of the disc. In addition, the parasagittal view of the lumbar spinal can reveal the conditions of the foraminal sections, which is helpful for assessing the conditions of foraminal stenosis, see Fig. 1.2(c). In clinical practice, parameters such as antero-posterior diameter, cross-sectional area of spinal canal on axial images and foraminal diameter on parasagittal images are typically used to quantify the severity of LSS [8]. However there is a lack of consensus in the literature and no diagnostic criteria are generally accepted [9]. As the pathologies exhibited in different areas are usually related, a more specific parametrisation and fitting of the structure, followed by a higher-level classification could contribute to more reliable, consistent and accurate diagnoses.

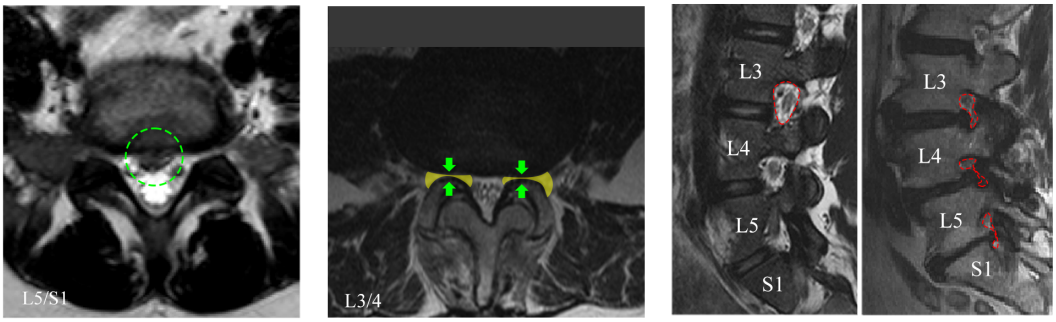
1.2 Methods and Contributions

We contribute in two main aspects, by proposing advanced parametric appearance models, and a weakly-supervised evidence pinpointing algorithm.



(a) (b) Anatomy

Figure 1.1: (a) Mid-sagittal view of a lumbar spine. Grey dashed lines show the raw axial scans. Red lines show the aligned disc-level planes, from which the axial images are extracted. (b) Anatomy of a L3/4 disc-level axial images.



(a) Central canal narrowing (b) Foreminal stenosis (c)

Figure 1.2: (a) A case with central canal narrowing caused by disc herniation. (b) A case with foraminal stenosis. The foramen is suppressed by the superior articular process and the disc (gree arrow). (c) Parasagittal scans of a normal case (left) and one with stenosis (right). Red circles outline the neural foramen.

1.2.1 Parametric appearance models

Parametric representation of an object category allows the leveraging of the prior knowledge by learning the statistics of the parameters in the population. The

training data usually consists of instances with landmarks annotated at consistent anatomical features. The appearance correspondence across the instances is built by aligning a deformable appearance (e.g., Active Appearance Model (AAM) [10]) or extracting local features at the landmarks([11, 12, 13]). The representations are often vectorised and used as inputs for training a classifier. During testing, the landmarks are detected in new, unseen instances, and the features are extracted and sent to the classifier for pathology classification. For a robust landmark detection, a prior model of the object class is learnt by formulating the statistics of the parameters, and the searching is conducted under the regularisation of the prior model. The deformable model is either holistic [10], which consists of the shape and aligned appearance, or part-based [14, 11, 12, 13], which represents an object by locally rigid parts with a shape capturing the spatial relationships among parts. In Deformable Part Model (DPM) the fitting process is implemented by local feature searching followed by a regularisation imposed through a prior model of the global shape. Various types of DPM instances have been proposed utilising advanced feature detection algorithms such as boosted regression [15], random forests [12], regularised mean-sift [14]; and shape optimisation methods such as pictorial structures [13] and non-parametric models [11]. However less attention has been paid to optimising the appearance representation in DPMs. We therefore focus on proposing advanced models of appearance delineation, in order to improve the precision of shape fitting as well as preserve the anatomical details which can be important in the context of medical imaging.

We introduce new appearance models referred to as Deformable Appearance Pyramids (DAPs) and Wavelet Appearance Pyramids (WAPs), built on Gaussian and wavelet image pyramids respectively. The object appearance is delineated by a multi-scale part-based model built on the image pyramid. The deformation is approximated by the translations of the parts as well as the linear appearance variations of the assembly of the parts. We introduce two methods to model the prior

and fit to new instances, one explicitly using a multi-variate Gaussian model and subspace LK algorithm [16], another implicitly using the Supervised Descent Method (SDM) [11].

1.2.2 Evidence pinpointing

To evaluate the pathological conditions based on radiological images, a clinician often inspects consistent and salient structures for localising the anatomies, then evaluates the appearance of certain local regions for evidence of pathology. In a computer-aided approach, by learning to localise these discriminative regions, or *pinpoint* the evidence, and describing them discriminatively could provide precise information for localising the anatomies and classifying pathology.

Conventionally, identifying and describing the regions of interest with respect to certain diseases requires strong supervision by experienced specialist with clinical background knowledge. To train a supervised method, the anatomical structures need to be localised and described by hand-crafted feature descriptors, which can be time consuming especially when dealing with large-scale data. We describe a method to automatically pinpoint the evidence regions as well as learn the discriminative descriptors in a weakly-supervised manner, i.e., only the class labels are used in training, and no other supervisory information is required. For localisation, we learn which features describe the anatomies saliently on a training set of aligned images. For classification, given the images with pathological labels, we learn the local features which provide evidence for discriminating between the normal and abnormal cases. We interpret evidence region pinpointing as a sparse descriptor learning problem [17, 18] in which the optimal feature descriptors are selected from a large candidate pool with various locations and sizes. Because of its large scale, the problem is formulated in a stochastic learning manner and the Regularised Dual Averaging algorithm [19, 20] is used for the optimisation.

The evidence pinpointing task is reminiscent of the multiple-instance problem

as described in [21] in which instances or features responsible for the classification are identified. Here, the learnt descriptors have several advantages over conventional hand-crafted representations, such as shape and appearance models, and local features, e.g., histogram of oriented gradient (HOG) [22] and local binary patterns [23]:

1. The training is weakly-supervised requiring no annotation of key features;

2. The learnt descriptors are more discriminative and informative, and therefore can contribute to better localisation and classification performance;

3. The evidence regions supporting the classification are automatically pinpointed which may be used by clinicians to determine the aetiology.

It is worth noting that the Convolutional Neural Network (CNN) architecture [24, 25, 26, 27] learns discriminative features from pathological labels with weak supervision as well, but requires large number of training samples and sufficient training. Instead of learning from raw image pixels, we formulate it as salient feature learning from a higher-level description of the image, which circumvents any need for the low-level feature training. As a result the optimisation is straightforward, consuming much less computing resource, and requiring no massive training data and no parameter tuning. Moreover, our descriptor learning method differs from the recent CNN based evidence pinpointing techniques [28, 29] in that we not only localise the evidence regions but at the same time give the description of these regions at optimal feature scales.

1.2.3 Research contributions

The main research contributions of this thesis are as follows:

1. A part-based parametric appearance model we refer to as a Deformable Appearance Pyramid. The parts are delineated by multi-scale Local Feature Pyramid (LFP) for superior spatial specificity and distinctiveness. A DAP

models the variability within a population with local translations of multi-scale parts and linear appearance variations of the assembly of the parts. We apply the DAP on the modelling of variability in patients with lumbar spinal stenosis and extend it to 3D for segmenting hip joints in 3D MR volumes.

2. Combining the wavelet image pyramid with DAP to improve the performance. The loglets [30] are adopted as the basis functions of the filter banks for their superior properties, such as uniform coverage of the spectrum (losslessness) and infinite number of vanishing moments (smoothness). The wavelet channels are complementary in the Fourier domain which enables the reconstruction of the appearance from a wavelet DAP.

3. An evidence pinpointing method to identify which specific regions and features of images contribute to a certain classification. The learning is weakly-supervised requiring only the pathological labels by clinicians and no other prior knowledge. It can also be applied to learn the salient description of an anatomy discriminative from background, in order to localise the anatomy before a classification step.

1.3 Clinical Data

1.3.1 Manual data preparation

In this thesis, we use data collected from routine clinics consistsing of T2-weighted MRI scans of over 600 patients with varied LSS symptoms. Each patient has anisotropic axial and sagittal scans with Digital Imaging and Communications in Medicine (DICOM) protocols. Due to the natural curvature of the lumbar spine, the axial scans are not aligned with the disc planes, therefore extracting the disc-level intervertebral images can be difficult. For this reason, we build the correspondence between the sagittal and axial scans, use the sagittal scans to lo-

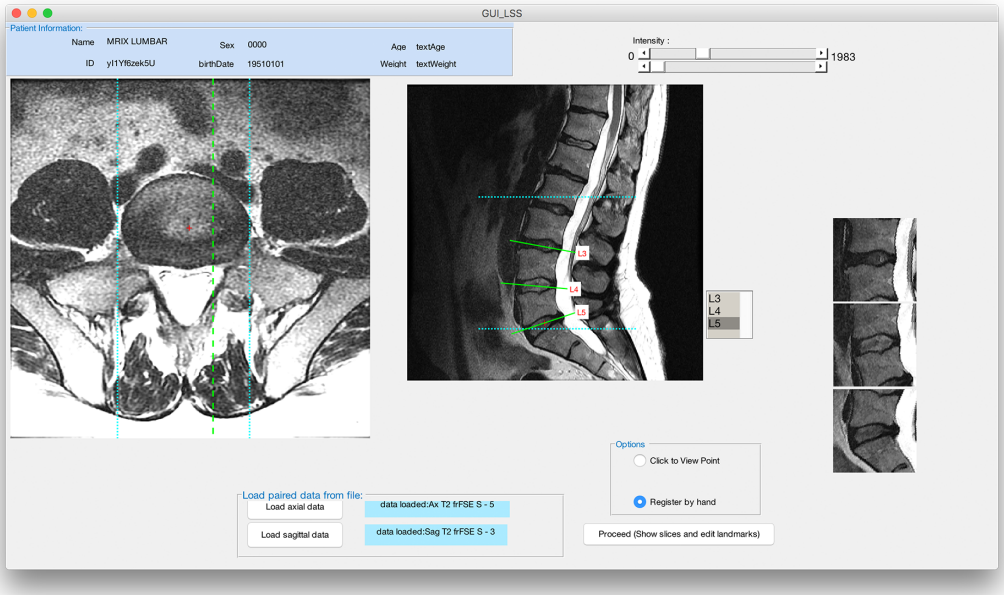
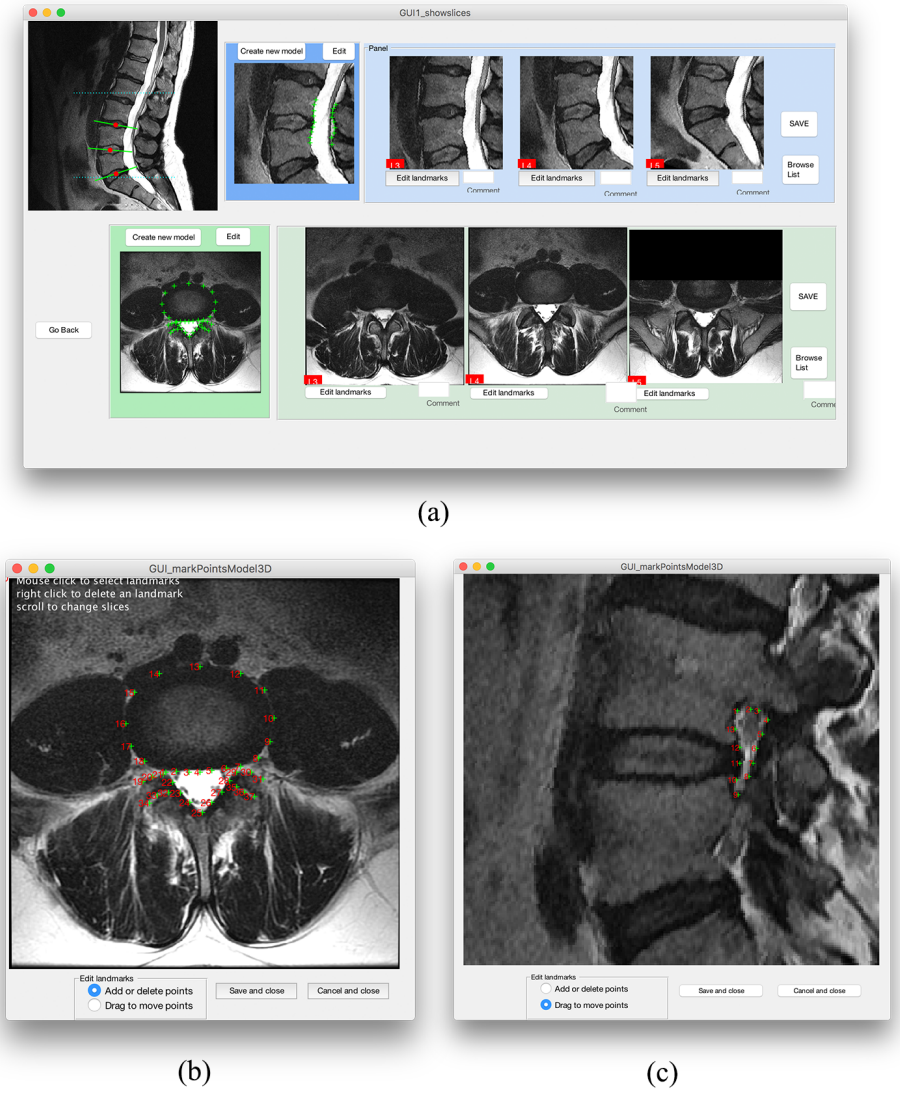


Figure 1.3: GUI to localise and align the discs, by hand.

calise and align disc-level planes by hand, and map the geometry to the axial scans to extract the disc-level images (Fig 1.3). Specifically, in both the sagittal and axial scans, the origin points and orientations of the images in the patient coordinate system are given in the dicom attributes `ImagePositionPatient` and `ImageOrientationPatient`, while the voxel spaces are given by the attributes `PixelSpacing` and `SpacingBetweenSlices`. Taking the sagittal scans as an example, the `ImagePositionPatient` gives the origin point $\boldsymbol{x}^0_{sag}$, while the `ImageOrientationPatient` gives the transformation matrix T_{sag}. Representing the voxel space with a matrix form $S_{sag} = \text{diag}\{\text{spaceX}, \text{spaceY}, \text{spaceZ}\}$ with each diagonal element being the voxel space in one of the three dimensions, the voxel location in the patient coordinate system is calculated by,

$$\boldsymbol{x}_{patient} = T_{sag}(S_{sag} \cdot \boldsymbol{x}_{sag}) + \boldsymbol{x}^0_{sag}. \tag{1.1}$$



(a)

(b) (c)

Figure 1.4: (a) GUI for extracting the sagittal and axial slicesby hand. (b)(c) GUI for annotating the axial and parasagittal images with landmarks.

Similarly the voxel coordinate in the paired axial scans can be calculated by,

$$\boldsymbol{x}_{axial} = S_{axial}^{-1} T_{axial}^{-1} (\boldsymbol{x}_{patient} - \boldsymbol{x}_{axial}^{0}), \tag{1.2}$$

in which S_{axial}, T_{axial} and $\boldsymbol{x}_{axial}^{0}$ denote the voxel space, transformation matrix and origin point of the axial scans respectively. In this way the correspondence between voxels in the sagittal and axial scans can be calculated.

We develop a MATLAB Graphical User Interface (GUI) to extract and annotate the slices from MRI scans. To obtain the precise disc-level axial images, in the sagittal scans, the intervertebral discs are localised and aligned (Fig. 1.3). The geometry is mapped to the axial scans through the patient coordinates, and the voxels are sampled at the corresponding locations to extract the aligned images (Fig. 1.4(a)). The axial planes are shifted along the spinal axis to inspect the pathology. The landmarks are then annotated on the images highlighting the features of interest (Fig. 1.4(b)). The parasagittal slices in the sagittal scans are also inspected, and the foramen images extracted and annotated (Fig. 1.4(c)). The workflow of preparing data with our interface is given in Fig. 1.5.

1.3.2 Dataset

The L3/4, L4/5, L5/S1 intervertebral images are localised and extracted using the developed GUI. We obtain three sets of 600 disc-level axial images from the three intervertebral planes respectively. Due to the different parameter settings The images are resampled to have an pixel space of 0.5 mm. All cases are inspected and annotated with classification labels with respect to the central stenosis and foraminal narrowing. The condition of central stenosis is classified with three grades, with grade one indicating normal/mild condition, grade two being moderate stenosis, and grade three being severe stenosis. The condition of foraminal narrowing is classified with two-class labels, with zero (or negative) indicating 'normal', and one (or positive) indicating 'abnormal'.

In addition the landmarks are available for the first 192, 198, 192 images in the three subsets, in which each image is delineated with 37 points outlining the disc, central canal and facet, see Fig. 1.4(b). In summary the dataset for validation contains three sets of 600 data with classification labels and three subsets of 192, 198, 192 data with labels and landmarks, which are referred to as weakly and densely annotated dataset respectively. In addition, from the sagittal scans of 200 cases, we

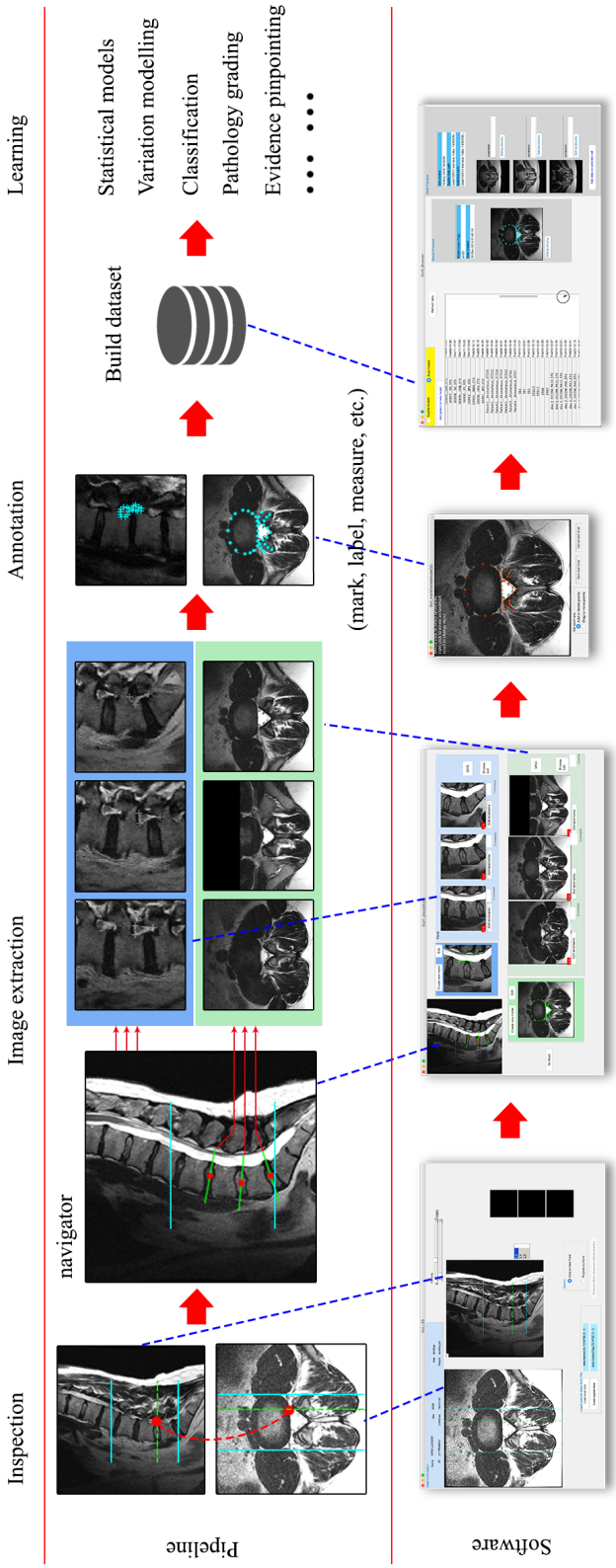


Figure 1.5: The workflow of LSS data preparation.

manually select the parasagittal planes where the nerve foramen can be located and extract 400 parasagittal images (200 on each side of the spines) around L3/4, L4/5, L5/S1 nerve foramina respectively. The contour of each foramen is annotated by 13 landmarks, see Fig. 1.4(c). Information of all the datasets are summarised in Table 1.1.

Table 1.1: Datasets configuration.

Dataset	L3/4	L4/5	L5/S1	Annotation
Densely annotated axial	192	198	192	landmarks and labels
Weakly annotated axial	600	600	600	labels only
Parasagittal	400	400	400	landmarks and labels

1.4 Thesis Outline

The rest of the thesis is organised as follows:

Chapter 2 reviews the methods in literature related to our work. The approaches and challenges of object appearance modelling are introduced in section 1. The pathology classification techniques are reviewed in section 2. The weakly supervised learning and evidence pinpointing methods are reviewed in section 3.

In chapter 3 we propose Deformable Appearance Pyramids for appearance representation and anatomy localisation. We detail how DAPs are constructed, trained and fitted and demonstrate that the appearance of an object can be delineated with multi-scale parts and that an associated deformation can be approximated by a set of locally rigid transformations of the parts. We derive an efficient fitting algorithm based on the weighted LK method and a regulariser utilising multi-scale landmark candidates. Both 2D and 3D DAP instances are validated.

In chapter 4 we propose the combination of wavelets and DAP for an improved appearance model we refer to as a Wavelet Appearance Pyramid. The object is decomposed into multi-scale complementary textures and each scale is further

decomposed into simpler parts. To achieve an explicit scale decomposition, filter banks are designed and arranged directly in the Fourier domain. The loglets [30] signal representation is adopted as the basis functions of the filter banks for their superior properties. The WAPs are applied for landmark detection and pathology classification with improved performance being observed in both tasks.

In chapter 5 we apply the loglet Local Feature Pyramids for part description in the Deformable Part Model framework, and validate the performance in the face alignment scenario. The proposed part descriptor improves the DPM by a significant margin, and gives better or competitive performance on the popular face datasets compared with the state-of-the-art methods in literature.

To be able to further learn on large scale weakly annotated data, in chapter 6, a weakly supervised learning approach is introduced to utilise data when no landmarks are annotated on the images. We describe a method to automatically pinpoint the evidence regions as well as learn the discriminative descriptors in a weakly-supervised manner, i.e., only the class labels are used in training, and no other supervisory information is required. The proposed weakly supervised methods show better or comparable performance when compared to conventional strongly supervised methods.

Chapter 7 summarises and discusses the methods presented, concludes the thesis and proposes some future research direction.

CHAPTER 2

Literature Review

In this chapter we review approaches and algorithms related to the tasks of appearance modelling, landmark detection, pathology classification and evidence pinpointing.

2.1 Deformable Models

2.1.1 Holistic and part-based models

Representation and segmentation of anatomical objects is of vital importance in the understanding of medical images. A standard approach which has proven robust and efficient, is to learn and leverage prior knowledge of the object garnered from statistics of its parametric form. To achieve this, the following steps are implemented: delineating the object class with a coherent parametric form; learning a prior model of the object class by formulating the statistics of the parameters; and fitting the parametric model to new, unseen instances while regularising the solution with the learned prior model.

The most commonly used strategy is to describe the objects holistically with shapes [31, 32, 33, 34, 35, 36, 37, 38] or deformable appearances such as morphable

models [39], statistical deformable models [40], and AAMs [10, 41, 42, 43]. The correspondence in the training data is established by annotating the landmarks at consistent features of interest from subjects. The prior knowledge is then usually learned through a linear model by applying eigen analysis, e.g. Principal Component Analysis (PCA). As a generative method, AAMs can not only achieve a robust segmentation, but also synthesise new instances and encode the appearance with compact parameters for higher-level interpretation, such as for the diagnosis and grading of pathologies. AAMs are widely adopted and have proven successful, but in clinical applications face challenges such as their sensitivity to local minima during fitting, and computational costs when built on 3D data.

In addition to the holistic methods, part-based models have shown superior performance in computer vision tasks including object detection and tracking. In part-based models an object is decomposed into locally rigid parts with a geometric model capturing spatial relationships among parts. Notable examples are sub-model AAMs [44, 45], Deformable Part Models (DPMs) [46, 47, 48], Constrained Local Models (CLMs) [49, 50, 51] and mixture-of-trees models [52]. Among these the models reported applied for clinical applications are sub-model AAMs and CLMs. For example in [50] the CLMs show superior performance over AAMs on brain and dental images. In [12] combined with random forests regression CLMs are reported to have the best performance in segmenting femur radiographs. The fitting process is implemented by local feature searching followed by a regularisation imposed through a prior model of the global shape. CLMs decompose the complex appearance into parts with simpler structures therefore suffer less from the high-dimension low-sample space problem when compared to AAMs. Moreover they are able to utilise advanced feature detection algorithms such as boosted regression [15], random forests [12], regularised mean-sift [14], and shape optimisation methods such as pictorial structures [13] and non-parametric models [11].

2.1.2 Challenges

The range of object representation and active fitting methods proposed in the literature strive to improve performance and precision. The methods have thus been adapted in various ways: to allow the prior models to compactly capture variation yet be able fit to unseen instances containing pathology; and prevent the fitting becoming trapped in local minima whilst maintaining a simplicity in object parametrisation and efficiency in fitting. We consider the challenge of local minima during fitting and how the choice of delineation (parametrisation) of objects can resolve this problem, but also result in a more flexible parts model which is, at the same time, efficient.

Local minima. Local minima are a problem facing all shape and appearance based methods. With the local minima the model can be trapped and fail to converge when initialised far away from the true location. This not only reduces the convergence range, which affects the range of initialisation, but also introduces large errors to the fitting results when a local minimum is identified as the global minimum. In both holistic and part-based methods, a coarse-to-fine strategy is often employed, which naturally increases the 'capture range' of the initialisation. However, even if at the finest level the model is close to the desired solution, the occurrence of local minima is still likely to divert the model from it [53].

Part-based models such as CLMs are plagued by the local minima problem due to their small local support and the large appearance variation in the dataset. The most effective strategy is to manipulate the scale. For instance, an efficient constrained mean shift method is proposed by [14, 51], in which a varying Kernel Density Estimate (KDE) is applied to perform coarse-to-fine fitting. The method starts with a smooth unimodal Gaussian model, and refines the fidelity by reducing the smoothness and increasing the number of modes. In [54] search for the local patches is conducted with coarse-to-fine resolution and the results are then

used as an initialisation for the AAM fitting. In [55] a hierarchy of shape models is designed to extend the CLM where the relationships between landmarks at each level is modelled by a MRF: the local models 'select' the best candidate points and the global model acts as a regulariser. They demonstrate an improvement in performance over CLMs. Despite the optimisation in feature searching algorithms, the choice of the feature scale (size of the image patches) itself is a trade-off between the location specificity and textural properties. Also the features at different landmarks themselves can have salient edges at varying scales, therefore a unitary scale for the descriptors for all landmarks will not capture faithfully all the salient features. In this thesis, we introduce a new descriptor called a Local Feature Pyramid (LFP) combining multi-scale local features at each landmark which gives a more comprehensive description. The shape fitting utilising multiple landmark estimations by the LFPs shows an ability to resist local minima.

Object class representation. In medical images, structural degeneration is often seen as local appearance changes. For example in MRIs of patients with LSS, vertebral degeneration is often seen as an abnormal shape along with local intensity changes which could indicate facet joint thickening and/or disc herniation and occasionally inflammation or fractures. In this instance, because the intensity and structural variations are related and coupled, a combined parametric delineation of shape and appearance might therefore offer a more robust segmentation. Representative methods using combined model are AAMs, atlas [56, 57] and CLMs.

AAMs have proven successful, but face challenges in the context of medical image analysis because:

1. AAMs model the interior region of the shape mesh, but for organs with convex shapes, a large proportion of textures of the interior region offers limited information while consuming a majority of the computational resources. Instead, there can be richer information nearby to the landmarks, at the periphery

of an organ boundary. Modelling the neighbourhood background can remedy this problem [58] but with an additional computational burden.

2. The memory usage and computational cost increases significantly when modelling volumetric data. The efficiency is reduced by the image warping process when performing the deformation, which is both expensive and complex to implement. Although there have been attempts to improve the tractability of 3D AAMs [59, 60], they have to either endure larger memory usage and slower speed, or sacrifice the precision by subsampling the data.

Atlas-based methods typically represent the training dataset with one or more atlas images. Comparing to AAMs, they require no well defined relation between regions and pixel intensities. Atlases have broad application in medical image segmentation and registration and are often used in computer aided diagnosis to measure the shape of an object or detect the morphological differences between patient groups. Various techniques for atlas construction are developed, especially for the brain images [61, 62, 40, 63]. However a deficiency is that it can be difficult to register the atlas to image instances when there are large shape and appearance variations in the dataset caused by pathologies.

In contrast to a holistic approach, CLMs describe the object with an assembly of local parts (patches) at key features. The parts are assumed to be conditionally independent of one another, an assumption that has demonstrated superior performance in computation and generalisation. This form of delineation readily allows integration with advanced feature searching techniques [51], and shape optimisation methods, e.g., a Bayesian inference [50] or density estimation [64]. However a deficiency is that as a coarse delineation, none of current methods give consideration to unbiasedly utilising, encoding and reconstructing the entire object appearance. We therefore introduce a novel part-based appearance model which can enhance the robustness and precision but also parametrise the whole appearance for subsequent

classification tasks such as diagnosis and grading.

Our approach is to start with a part-based model, by parametrising objects as an assembly of object parts, but with the parts being multi-scale local appearance captured by a LFP. This multi-scale approach we call the Deformable Appearance Pyramid (DAP) overcomes problems of local minima when searching for landmark locations, and the pyramid structure allows the appearance model to fully cover the object interiors and capture the landmark context, allowing the resulting DAP to have generative capabilities. The part-based form also gives us flexibility in our choice of fitting strategy.

2.2 Disease Classification

Disease classification based on radiological images is a key task in Computer-Aided Diagnosis (CAD). Advances in medical imaging technology and machine learning have greatly enhanced interpretation of medical images, and contributed to early diagnosis. The development of CAD systems to assist physicians in making better decisions has been the area of interest in the recent past. Disease classification aims to provide predictions as a second opinion in order to assist physicians in the detection of abnormalities and quantification of disease progress.

2.2.1 Feature extraction

In a classical classifier, each object used for training and testing is represented by a feature vector extract from the image, and a discrimination rule is applied to classify a given test vector. The most commonly used feature extraction techniques can be divided into two general categories, namely hand-crafted features and unsupervised feature based approaches. Broadly speaking, hand-crafted features refer to those which can be connected to specific measurable attributes in the image and have some degree of interpretability. Unsupervised or weakly supervised feature approaches,

such as discriminative descriptor learning [18] and deep learning based methods [25] require less human interference and learn the optimal descriptors from large numbers of training examples to characterise and model image appearance. In this type of learning, the actual features generally remain latent (or hidden).

Hand-crafted features often describe the anatomical structures with shapes, local textures or a combination of shape and texture. Texture descriptors are commonly used in verious types of medical images. For instance in [65] a series of wavelet and tissue texture features are used for automated Gleason grading of prostate pathology images. In [66], Fisher vector descriptors are extracted from the regions of interest for the classification of Parkinson's disease from diffusion MRI data. Otalora et al. [67] propose to represent and encode the tissue images with Riesz wavelet to identify anaplastic medulloblastoma in histopathology images. An anatomy can also be represented by the combination of shape and local features. For example the hierarchical shapes and local features such as local binary patterns and Gabor wavelets has been used to describe the facial dysmorphology of down syndrome [23]. Similarly in [68], it is shown that the combined descriptors of shapes and local features could result in accurate breast cancer grading.

Supervised and manually designed feature descriptors are usually based on landmarks therefore require expensive human labour and rely on expert knowledge. This motivates the design of efficient feature learning techniques to automate and generalise the design of feature descriptors. In addition, the increased data scale and the availability of more powerful computing sources enable the efficient learning and optimisation of descriptors, resulting in superior classification performance. For example in [69], the spatio-temporal features are learnt with a recurrent neural network, which incorporates a deep hierarchical visual feature extractor and a temporal sequence learning model. A joint learning framework is proposed with knowledge transfer across multi-tasks. A tensor-based feature learning approach is proposed in [70] for whole-brain fMRI classification. In [71], an unsupervised feature

learning framework is proposed for automatic diagnosis of ovarian carcinomas. It is composed of a sparse tissue representation and a discriminative feature encoding scheme based on a hybrid model.

2.2.2 Classifiers

After feature extraction, the next important step is building, training and validating the classifier. Various types of classifiers have been applied to computer-aided diagnosis to deal with the diversity in different clinical tasks and contexts. Notable methods are Support Vector Machine (SVM), decision trees, and more recently CNNs. For example Wong et al. [72] combine the SVM classifier with a biomechanical model to identify the infarction from cardiac CT images. In [71], a linear multiclass SVM classifier and the ability to diagnose whole slide ovarian carcinoma images is demonstrated. A multi-output decision tree regressor is proposed by Chandran et al. [73] to predict trabecular bone anisotropy from clinical quantitative CT images. Similarly a regularised tree boosting algorithm is presented by Li et al. [74] and combined with multiple instance learning for colorectal cancer detection. In [75], the features are identified with CNN, and classified using a Random Forest to quantify coronary artery calcification in cardiac CT angiography. Other classifiers have also been adopted and developed to suit the specific clinical context. For instance an evidential K-Nearest-Neighbour classifier is developed by Lian et al. [76] to predict the outcome in cancer therapy. A sparsity conducted classification framework is introduced in [77] to classify the colon cancer and lung cancer. The sparse representation is adopted to reduce the feature dimension. To classifying the disentangling disease heterogeneity, a nonlinear algorithm is developed by Varol et al. [78] to learn the integrated binary classification and subpopulation clustering.

2.3 Weakly Supervised Learning and Evidence Pinpointing

2.3.1 Weakly supervised learning

Weakly supervised learning methods seek to extract the useful information from a dataset with overview meta information such as the classification labels but without requiring any precise annotations such as landmarks. They are favourable especially when learning on large scale problems because the dense annotation can be infeasible on large amount of data.

One of the popular approaches suitable for weakly supervised learning is the deep convolutional neural networks (CNNs). A CNN is a type of feed-forward artificial neural network in which the connectivity pattern between its neurons is inspired by the organization of the animal visual cortex. There are typically three major techniques that successfully employ CNNs to medical image classification [25]: (1) training the CNN from scratch [79, 80, 24, 81, 75]; (2) using 'off-the-shelf CNN' features (without retraining the CNN) as complementary information channels to existing hand-crafted image features [82, 83, 84]; and (3) performing unsupervised unsupervised pretraining on natural or medical images and fine-tuning on medical target images using CNN or other types of deep learning models [84, 85, 86, 87].

CNN architectures have updated the benchmarks in many clinical tasks, but require large number of training samples and sufficient training. One of the reasons is that they learn the features from raw image pixels, which usually require several layers to reach an advanced feature level. Instead of learning from raw image pixels, here we propose a weakly supervised learning method [88] formulating the problem as salient feature learning from a higher-level description of the image, which circumvents the low-level feature training. As a result the optimisation is straightforward consuming much less computing resource, and requiring no massive

training data and no parameter tuning.

2.3.2 Evidence pinpointing

A clinician often inspects the appearance of certain local regions for evidence of pathology. In conventional computer-aided diagnosis, the computer can predict the pathological labels, but the question of what local features, or evidence is supporting the prediction still needs to be answered. Learning to automatically highlight, or pinpointing the relevant regions can not only show the evidence supporting the conclusion, but also reveal the evidence that may have been overlooked or unexpected by the clinician. The results could therefore provide additional information for a better understanding of a certain disease.

The task scenario is reminiscent of Multiple-Instance Learning (MIL) [21, 89]. MIL was first introduced in [90] in the context of drug activity prediction. A multiple-instance problem involves ambiguous training examples: a single example is represented by several feature vectors (instances), some of which may be responsible for the observed classification of the example; yet, the training label is only attached to the example instead of the instances. Far beyond the drug activity prediction problem, the multiple-instance problem emerges naturally in a variety of challenging learning problems in computer vision, including natural scene classification, content-based image retrieval, image categorization and object detection and recognition. Generally speaking, the goal of all these problems is to learn visual concepts form labelled images. For example, the approach in [21] identifies instances that are relevant to the observed classification by embedding bags into an instance-based feature space and selecting the most important features. A similarity measure between a bag and an instance is defined. The coordinates of a given bag in the feature space represent the bags similarities to various instances in the training set. The embedding produces a potentially high dimensional space when the number of instances in the training set is large. A subset of most relevant fea-

tures are selected because many features may be redundant or irrelevant as some of the instances might not be responsible for the observed classification of the bags, or might be similar to each other. A joint approach is chosen that constructs classifiers and selects important features simultaneously. Since each feature is defined by an instance, instance selection is essentially feature selection.

In CNNs, the task is similar to saliency map computation. For example in [91] the convolutional network is visualised by the class saliency map, trained on the large-scale ImageNet challenge dataset. The procedure is related to the ConvNet training procedure, where the back-propagation is used to optimise the layer weights. In [28], the CNN is trained to not only predict the class labels but also localise the objects with image-level supervision. In [25], the approaches of visualising disease regions in the medical imaging scenario are investigated. More recently the method in [91] is adopted and modified to produce a heat-map which pinpoints the regions in the image responsible for the prediction [29]. The map lights up pathological areas of the prediction specific to the trained task in the image.

In this thesis we propose a learning method to identify which specific regions and features of images contribute to a certain classification. We formulate evidence pinpointing as a sparse descriptor learning problem [17, 18]. Because of the large computational complexity, the objective function is composed in a stochastic way and is optimised by the Regularised Dual Averaging algorithm [20]. We go on to demonstrate that learnt feature descriptors contain more specific and better discriminative information than hand-crafted descriptors contributing to superior performance for the tasks of anatomy localisation and pathology classification respectively. Our descriptor learning method differs from the above mentioned CNN-based techniques [28, 29] in that we not only localise the evidence regions but at the same time give the description of these regions at optimal feature scales.

2.4 Other Related Works

2.4.1 Loglets

The idea of designing and tiling filters in the Fourier domain has led to efficient decomposition of local structures at multiple resolutions and orientations, e.g., steerable pyramids [92], Gabor filters [93, 94, 95], log-Gabor filters [96, 97], curvelets [98], contourlets [99], loglets [30], to name but a few. A Gabor function is a complex oscillation multiplied by a Gaussian envelope and in the Fourier domain manifests as a Gaussian function shifted away from the origin. A log-Gabor filter is a Gaussian on a logarithmic frequency scale, which has a wider bandwidth towards the higher frequencies and leads to a compact form under scaling transformations when compared with Gabor filters. A generalisation to the log-Gabor function is the loglets, as proposed by Knutsson [30], with enhanced properties such as a uniform coverage of the spectrum and an infinite number of vanishing moments (smoothness). Loglets have invariance to illumination, but because they are invariant also to sample shift they suffer less information loss in sampling caused by the limited resolution of discrete images.

2.4.2 Feature-to-shape regression

In discriminative shape fitting approaches the problem can be formulated as feature-to-shape regression, which learns a mapping from image features to landmark locations. For example in [10] the AAMs are by learning a linear regression between the increment of motion parameters and the appearance differences. The linear regressor is a numerical approximation of the Jacobian [10]. Following this idea, several discriminative methods that learn a mapping from appearance to shape increments have been proposed. Gradient Boosting, first introduced in [100], has become one of the most popular regressors because of its efficiency and the ability to model nonlinearities. In [101, 102], it is shown that using boosted regression for

AAM discriminative fitting significantly improved over the original linear formulation. In [103], the pose indexed features are incorporated to the boosting framework, where not only a new weak regressor is learned at each iteration but also the features are re-computed at the latest estimate of the landmark location. Beyond the gradient boosting, in [104], the kernel regression is explored to map from image features directly to landmark location achieving surprising results for low-resolution images. Recently, Cootes et al. [105] investigated Random Forest regressors in the context of face alignment. At the same time, Sanchez et al. [106] proposed to learn a regression model in the continuous domain to efficiently and uniformly sample the motion space. In the context of tracking, Zimmermann et al. [107] learned a set of independent linear predictors for different local motion and then a subset of them is chosen during tracking. More recently in [11, 108], a Supervised Descent Method (SDM) is proposed which learns the generic descent directions of featuretnete-to-shape regression in a supervised manner. During training, the SDM learns a sequence of descent directions that minimizes the mean of non-linear Least Squares functions sampled at different points. In testing, SDM minimises the objective function using the learned descent directions without computing the Jacobian nor the Hessian matrix. In their work the benefits of the SDM approach is validated in the facial landmark detection and face tracking context.

2.5 Summary

In this chapter we reviewed the strategies of modelling the object appearance, and addressed the challenges faced when applying the appearance models to the task of landmark detection. Several feature extraction techniques and classifiers were reviewed for the task of pathology classification. In addition, we introduced the problem of evidence pinpointing and reviewed related weakly-supervised methods. In the next chapter, we introduce our first major contribution which is a new method

for object appearance modelling.

CHAPTER 3

Deformable Appearance Pyramids

for Appearance Modelling and Landmark Detection

Object class representation is of vital importance for medical image analysis tasks such as modelling the appearance variations and localising anatomical features. Parametric representation of an object category allows the leveraging of the prior knowledge by learning the statistics of the parameters in the population. The representations are often vectorised and used as inputs for training a localiser (Fig. 3.1). The training data usually consists of instances with landmarks annotated at consistent anatomical features. For a robust landmark detection on the testing data, a prior model of the object class is learnt by formulating the statistics of the parameters, and the searching is conducted under the regularisation of the prior model.

In this chapter, we introduce a new part-based appearance model we refer to as a Deformable Appearance Pyramid [109, 110]. We start by deducing the Gaussian shape prior with density estimation theory [111] in section 3.1. The DAP representation is introduce in section 3.2. We present a fitting method based on subspace Lucas-Kanade (LK) algorithm in section 3.3. Section 3.4 validates the

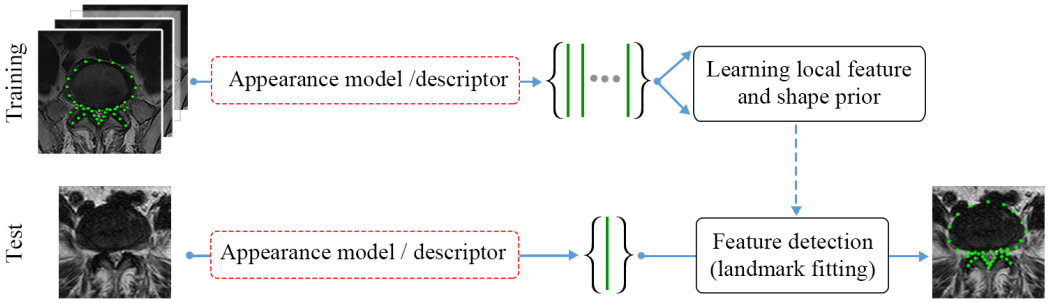


Figure 3.1: Pipeline of landmark detection with appearance models.

performance of DAP instances in landmark detection on the problem of LSS. Section 3.5 validates the performance in landmark detection in 3D, on the problem of hip impingement.

3.1 Shape Prior Modelling with Density Estimation Theory

Shape prior modelling is to model and learn the statistics of shapes of an object class given the training samples, and using it to regularise the shape when fitting to a new object instance. Assuming a multi-variable Gaussian distribution of the shapes, with density estimation theory [111], the statistic of the shapes can be characterised by the Mahalanobis distance from the mean, with the variances given by the eigenvalues $\{\lambda_i\}$ obtained from PCA. A more sufficient statistics and efficient calculation of this model can be obtained by a piecewise estimator of the eigenvalues *spectrum*, by which the Mahalanobis distance is decomposed into a much lower dimension Mahalanobis distance in the eigenspace (a subspace spanned by the eigenvectors) and an Euclidean distance in the orthogonal space [111]. We deduce the shape prior modelling with density estimation theory in detail as follows.

A shape in 2D can be described by a point distribution model denoted by $\boldsymbol{s} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N] \in \mathbb{R}^{2N}$, in which N is the number of landmarks, and $\boldsymbol{x}_i$ is the

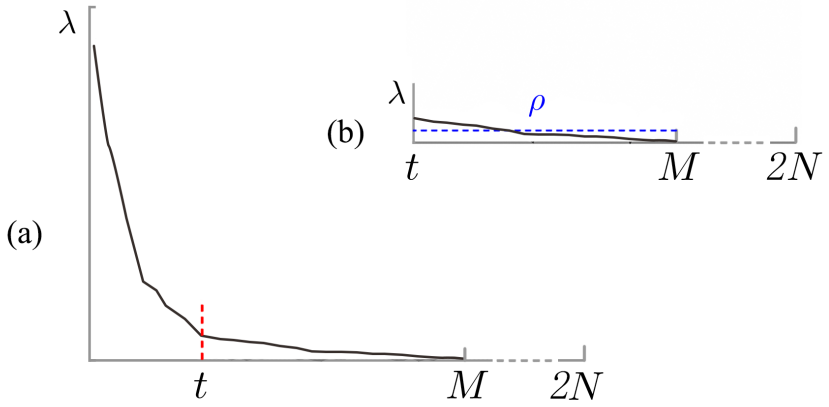


Figure 3.2: (a) A typical eigenvalue spectrum by applying PCA on the covariance matrix of the shapes. The spectrum can be described piecewise in three ranges: $[0, t]$, $[t, M]$, and $[M, N]$. (b) Approximate the eigenvalues in range $[t, M]$ with a constant value ρ.

coordinate of the i-th landmark, i.e, $\boldsymbol{x}_i = [x_i, y_i]$. Given a set of training shapes $\{\boldsymbol{s}^k\}_{k=1}^M$ from an object class Ω, we wish to parameterise the prior knowledge of the shapes by calculating the probability (density) of an instance $\boldsymbol{s}$ in the shape assembly, i.e., $p(\boldsymbol{s}|\Omega)$. Assuming a Gaussian distribution of the shape variations, the likelihood of $\boldsymbol{s}$ belonging to the shape class is given by,

$$p(\boldsymbol{s}|\Omega) \propto \exp(d(\boldsymbol{s}, \bar{\boldsymbol{s}})) = \exp\left(-\frac{1}{2}(\boldsymbol{s} - \bar{\boldsymbol{s}})^T \Sigma^{-1} (\boldsymbol{s} - \bar{\boldsymbol{s}})\right), \tag{3.1}$$

where $\Sigma \in \mathbb{R}^{2N \times 2N}$ is the covariance matrix and $\bar{\boldsymbol{s}}$ is the mean shape. Calculating the likelihood using (3.1) directly is infeasible as Σ is of high dimension. A more efficient way is to reduce the dimension first. Note that Σ is typically low-rank and has no more than M non-zero eigenvalues, therefore a lossless dimension reduction is available. A standard method is to apply PCA on Σ, by which we can obtain a diagonal matrix of the first M eigenvalues $\Lambda = \text{diag}\{\lambda_1, ..., \lambda_M\} \in \mathbb{R}^{M \times M}$, arranged in descending order of magnitude, and the corresponding eigenvector matrix $P \in \mathbb{R}^{2N \times M}$ which spans a subspace. A typical eigenvalue spectrum is shown in

Fig 3.2(a). Accordingly the covariance matrix can be diagonalised as,

$$\Lambda = P^T \Sigma P, \tag{3.2}$$

in which the covariance is decoupled. Substituting (3.2) into (3.1) the Mahalanobis distance becomes,

$$\begin{aligned} d(\boldsymbol{s}) =& (\boldsymbol{s} - \bar{\boldsymbol{s}})^T \Sigma^{-1} (\boldsymbol{s} - \bar{\boldsymbol{s}}) \\ =& (\boldsymbol{s} - \bar{\boldsymbol{s}})^T P^T \Lambda^{-1} P (\boldsymbol{s} - \bar{\boldsymbol{s}}) \\ =& \boldsymbol{b}^T \Lambda \boldsymbol{b} \\ =& \sum_{i=1}^{M} \frac{b_i^2}{\lambda_i}, \end{aligned} \tag{3.3}$$

where $\boldsymbol{b}$ is the projection of the mean-normalised shape onto the subspace spanned by P. In this way, the dimension of the problem has been reduced to M. However it can be still too high for an efficient computation because in order to obtain $\{b_i\}_{i=1}^{M}$ we need to calculate the projection of the shape onto M bases vectors. Note the fact that as the shapes are correlated, the eigenvalue spectrum usually decays rapidly and flattens out after a short range, as shown in Fig 3.2. This makes possible a precise approximation using a small number of t significant projections. A typical value of t is chosen by $t = \min\{t' | \sum_{i=1}^{t'} \lambda_i / \sum_{i=1}^{M} \lambda_i > 98\%\}$. The eigenvalue spectrum can be approximated as piecewise by,

$$\hat{\lambda}_i = \begin{cases} \lambda_i & i \in [1, t] \\ \rho & i \in (t, M] \\ 0 & i \in (M, 2N] \end{cases}, \tag{3.4}$$

in which ρ is the Maximum Likelihood (ML) estimator of the eigenvalues in the 'flat' range (Fig 3.2(b)), which can be proven to be the arithmetic average, $\rho =$

$\frac{1}{M-t}\sum_{i=t+1}^{M}\lambda_i$, see [111] for the detailed deduction. Note however that the value of ρ here is different from the one in [111], where it is the mean of $\{\lambda_i\}_{i=t+1}^{2N}$. The formulation here yields a more precise estimation because we have considered the fact that $\{\lambda_i\}_{i=M}^{2N}$ are always zeros, see Fig 3.2(b). Accordingly, the eigenspace is divided into a principal subspace $F = \{P_i\}_{i=1}^{t}$ and its orthogonal space $\bar{F} = \{P_i\}_{i=t+1}^{M}$ (P_i denotes the i-th eigenvector in P).

By the formulation in (3.4), we can decompose (3.3) into the Mahalanobis distance in F and a Euclidean distance in $\bar{F}$.

$$d(\boldsymbol{s}) = \sum_{i=1}^{t}\frac{b_i^2}{\lambda_i} + \frac{1}{\rho}\sum_{i=t+1}^{M} b_i^2. \tag{3.5}$$

The summation of the projections in the second term can be calculated efficiently by,

$$\sum_{i=t+1}^{M} b_i^2 = ||\boldsymbol{s} - \bar{\boldsymbol{s}}||^2 - \sum_{i=1}^{t} b_i^2. \tag{3.6}$$

Therefore we have,

$$d(\boldsymbol{s}) = \boldsymbol{b}^T\Lambda\boldsymbol{b} + \frac{1}{\rho}(||\boldsymbol{s} - \bar{\boldsymbol{s}}||^2 - ||\boldsymbol{b}||^2). \tag{3.7}$$

From now on we refer to $\boldsymbol{b} \in \mathbb{R}^t$ and $\Lambda \in \mathbb{R}^{t\times t}$ as the values of the first t components, instead of the first M, a reasonable abuse of the terminology. We also modify the notation $P \in \mathbb{R}^{2N\times t}$ as the matrix of the first t eigenvectors. $\boldsymbol{b}$ can be calculated by projecting the shape onto the subspace spanned by P,

$$\boldsymbol{b} = P^T(\boldsymbol{s} - \bar{\boldsymbol{s}}). \tag{3.8}$$

Replacing the Mahalanobis distance in (3.1) with (3.7), the shape prior attained is,

$$p(\boldsymbol{s}|\Omega) = p_F(\boldsymbol{s}|\Omega)p_{\bar{F}}(\boldsymbol{s}|\Omega) \propto \exp(\boldsymbol{b}^T\Lambda\boldsymbol{b})\exp\left(\frac{1}{\rho}(||\boldsymbol{s} - \bar{\boldsymbol{s}}||^2 - |||\boldsymbol{b}||^2\right), \tag{3.9}$$

which gives a computational efficient yet statistically sufficient solution of the likelihood of a shape instance in the shape class Ω.

3.2 DAP for Appearance Representation

In this section we introduce the Local Feature Pyramid (LFP) and the Deformable Appearance Pyramid representation.

3.2.1 Local feature pyramids

The local appearance at a landmark is typically described by an image patch at a certain scale. For sharper structures, a smaller scale can give a more precise pixel location. At blurry structures however, the scale should be large enough to cover distinguishable textural information. A good feature descriptor is expected to have a high spatial specificity (pixel location) while maintaining good distinctive ability (textural properties). Due to the uncertainty principle in signal processing, a single scale patch cannot achieve both. We therefore propose a multi-scale part descriptor, with the smaller scales containing local high frequency features, and the larger scales low frequency components.

An L-level LFP at a landmark consists of L patches centred there with increasing scales and decreasing resolutions at octave intervals. The first level patch is the smallest one with the finest resolution. A patch in the l-th level has l octaves larger scale and lower resolution, which keeps the same size in pixel across all levels, see the 2D and 3D examples in Fig. 3.3. The representation is reminiscent of a wavelets description in which to obtain high specificity in both location and frequency, the signal is expanded over a number of scales in octave intervals forming a joint time-frequency tiling [112].

A robust landmark searching can be implemented by performing the feature detection at individual scales and combining the results. The LFP at a landmark

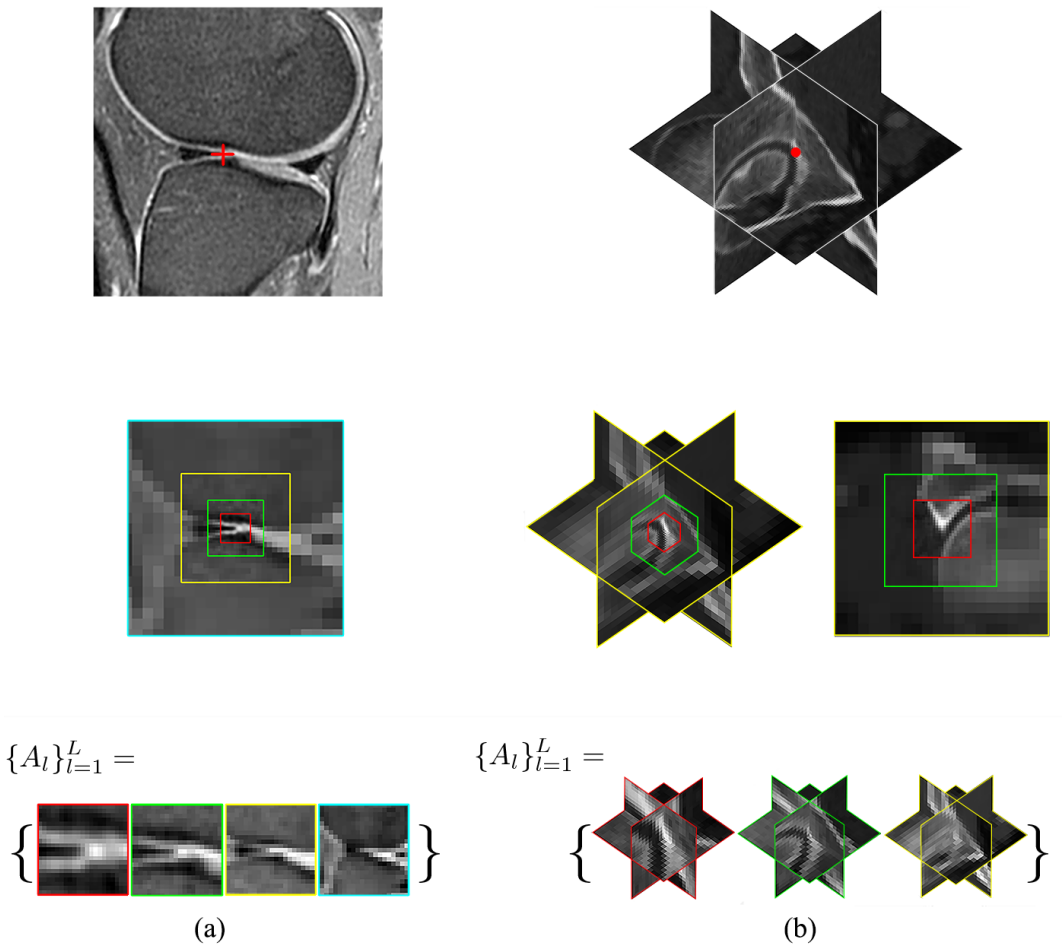


Figure 3.3: (a) 2D LFP; (b) 3D LFP. Top row: A landmark at an image instance; Middle row: The LFP at the landmark; Bottom row: 2D patches or 3D subcubes at all levels in a LFP are concatenated forming a profile of the local feature.

is denoted as $\{A_l\}_{l=1}^{L}$, with patch A_l giving the profile of the local feature at the l-th scale. Running feature searching at each scale we can obtain a response map which represents probabilistic distribution of the landmark location $p(\boldsymbol{x}|A_l)$. The response maps from four level profiles is illustrated in Fig. 3.4(b to e). The maps are reminiscent of the multi-scale products in [113] where the local maximum at multiple scales are combined for spot detection.

The probabilistic distribution of the landmark combing all the predictions in the LFP can be formulated as a product,

$$p(\boldsymbol{x}|\{A_l\}_{l=1}^{L}) \propto \prod_{l=1}^{L} p(\boldsymbol{x}|A_l). \tag{3.10}$$

An example of a product combination is shown in Fig. 3.4(f). We can see that the combined response map has a sharper peak at the true location, and the local minima are suppressed.

It is worth noting that multi-resolution and multi-scale techniques have been widely used in computer vision. For example in [114], the local feature is described with Scale-Invariant Feature Transform (SIFT) at different levels of detail, and in [115], a 'pooling' across adjacent scales is performed. In our feature descriptor all scales are combined in an LFP for a comprehensive local feature profile at individual landmarks, with the aim to enhance the robustness to local minima and feature saliency:

1. Resistance to local minima, see Fig. 3.4. Local feature detectors are plagued by the problem of ambiguity. This ambiguity is evident in the distribution of landmark locations (i.e., the response map) obtained from a feature detector, see Fig. 3.4(b). In [51], a multi-scale parametrisation of the response map is used to seek for the true position. Our feature pyramid however, deals with this problem from a different perspective: it calculates multi-scale response maps (see Fig. 3.4(b) to (e)) from multi-scale patches, and combines the responses

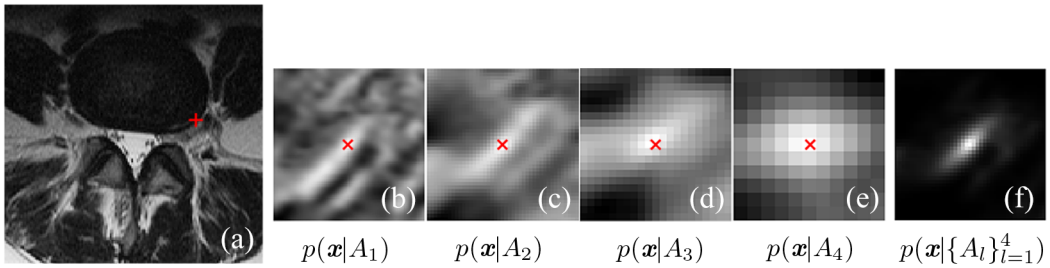


Figure 3.4: (a) A marked feature point (red cross); (b) to (e) Response maps from four level local features in a LFP at the landmark. Red crosses denote the true locations. The smaller scales are plagued by the problem of ambiguity, while the lager scales have low spatial specificity; (f) A product combination of the response maps, which enhances the specificity and suppresses the ambiguity.

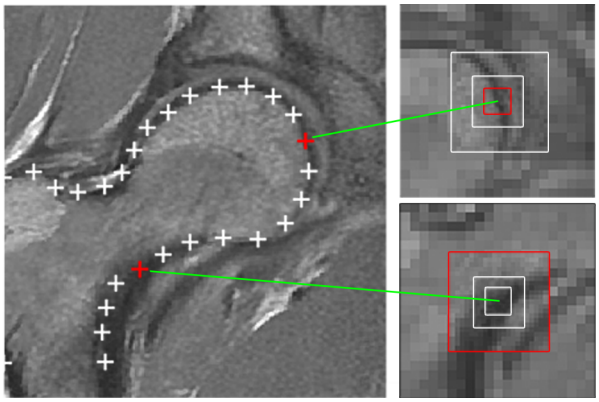


Figure 3.5: Local features are salient at certain scales. A LFP is able to preserve the salient scales (red rectangles).

to estimate the true position. The larger scale ensures a wider support range while the small scale yields a high precision.

2. Enhanced distinctive ability, see Fig. 3.5. At certain scales, local features at a landmark often show consistency across cases while being distinctive from the background. We refer to these scales as the 'salient scales'. Different measurements can be used for determinating the salient scales [116]. However as noted earlier, a single-scale descriptor will either be too small to capture the texture or too spread-out to give its precise location. In comparison, the feature pyramid can preserve the salient features at whatever scales they appear (e.g., the red patches in Fig. 3.5).

3.2.2 Deformable Appearance Pyramid

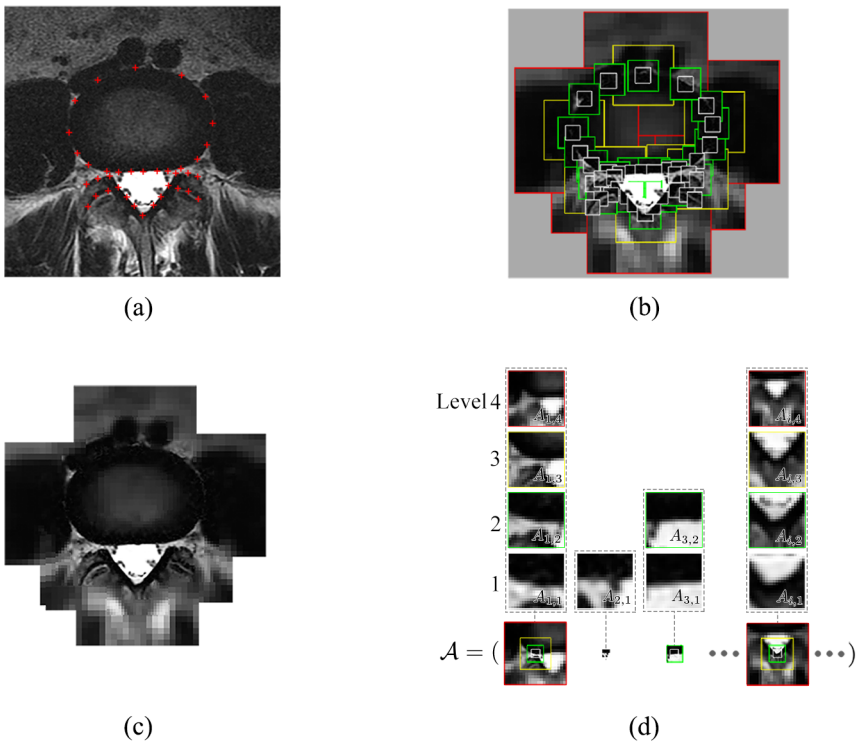


Figure 3.6: (a) Image with landmarks. (b) DAP with 4 level feature pyramids. (c) The DAP delineation. (d) Concatenated LFPs form a 1D DAP vector $\mathcal{A}$.

An DAP is a part-based model with each part delineated by a LFP. The DAP consists of two elements: $\{\mathcal{A}, \boldsymbol{s}\}$, with $\mathcal{A}$ being the *assembly* of the feature pyramids and $\boldsymbol{s}$ the shape. To reduce the overlap at coarser levels, we 'trim' the DAP and keep fewer patches at landmark intervals at the larger scales. The principle of trimming is to obtain an even coverage of the appearance at each level, see Fig. 3.6(b). In practice, a simple trimming algorithm can be designed to iteratively delete the landmark that has least distance from its neighbourhood until a distance criterion is matched. Alternatively a landmark to preserve can be selected by hand to highlight the anatomical features of interest. Denoting $\mathcal{K}_l$ as a subset of natural numbers $\{1, ..., N\}$ indicates the landmarks preserved at the l-th scale. The assembly of the trimmed parts can be denoted as $\mathcal{A} = \{\{A_{i,l}\}_{i \in \mathcal{K}_l}\}_{l=1}^{L}$, see Fig. 3.6(d) for an

example. The parts are sorted in same order across the dataset in $\mathcal{A}$ in order to keep the topology consistent. $\mathcal{A}$ is then flattened into a 1D vector serving as the profile of the whole object appearance.

Given the training set we can extract an $\mathcal{A}$ from each image and obtain a set of training data $\{\mathcal{A}_1, \mathcal{A}_2, ...\}$. By extracting the local features from the corresponding landmarks, the shape variation in the training set has already been removed and a better pixel-to-pixel correspondence achieved, therefore $\mathcal{A}$ can be viewed as 'shape-free' appearances and an extra image warping as necessary in AAMs is avoided. It should be noted that at larger scales, the structural deformation might be included. However this is acceptable because larger scales have lower resolution and therefore are less sensitive to the shape variations. $\mathcal{A}$ can be visualised by recovering the dimension and location of each feature patch, padding and placing smaller scale patches on top of larger ones, see Fig. 3.6(c). To obtain a statistical model of the shape-free appearance, we normalise the mean and variance of each $\mathcal{A}$ and apply PCA on the training samples. A new instance can be linearly modelled by,

$$\mathcal{A} = \bar{\mathcal{A}} + P_{\mathcal{A}} \boldsymbol{b}_{\mathcal{A}}, \tag{3.11}$$

in which $\bar{\mathcal{A}}$ is the mean, $P_{\mathcal{A}}$ spans the eigenspace and $\boldsymbol{b}_{\mathcal{A}}$ is the appearance parameters in the subspace.

3.3 Deformable Appearance Pyramid Fitting

The DAP is parametrised by the appearance of the assembly of parts as well as the shape capturing spatial relationships. It therefore fits and synthesises new instance by adjusting global appearance parameter $\boldsymbol{b}_{\mathcal{A}}$, and estimating local translations for individual patches with a regulariser imposed on the shape $\boldsymbol{s}$. We follow the two-step fitting strategy commonly used in part-based models, i.e, local feature searching followed by a geometrical regularisation.

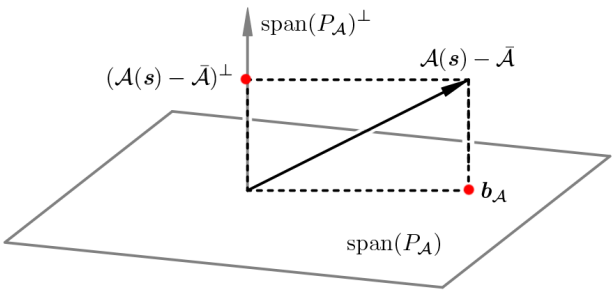


Figure 3.7: The eigenspace of appearance variations and its orthogonal space. Appearance is fitted in the eigenspace. LK based landmark estimation is performed in the orthogonal space, where the appearance variations are projected out.

3.3.1 LK based simultaneous local feature searching and appearance fitting

The LK algorithm attempts to find the parameters $\boldsymbol{p}$ to minimise the difference between a template T and a source image J,

$$\boldsymbol{p} = \arg\min ||J(\boldsymbol{p}) - T||^2, \tag{3.12}$$

where $\boldsymbol{p}$ can be image translation or warping, and $J(\boldsymbol{p})$ is the image after the transform. To enhance the robustness and efficiency respectively, two extensions have been made, namely weighted LK and inverse gradient descent [16]. The weighted LK can be posed as,

$$\boldsymbol{p} = \arg\min ||J(\boldsymbol{p}) - T||^2_Q, \tag{3.13}$$

where Q is the weighting matrix usually representing a linear transform such as a subspace projection in the AAMs [41], weighted subspace projection [117], or Gabor filtering in the Fourier LK [118, 43].

We derive a subspace LK for the DAP fitting, with a further simplification by applying the conditional independence assumption of the part-based models. Specifically, the difference between the template and the textures it covers can be caused by the appearance variation of the object and the departure of the model

from the true position. Accordingly they can be dealt with in two subspaces: the eigen-space $span(P_{\mathcal{A}})$ accounting for the appearance variation and its orthogonal space $span(P_{\mathcal{A}})^{\perp}$ to predict the landmark shift. The appearance parameters $\boldsymbol{b}_{\mathcal{A}}$ can be calculated by projecting $\mathcal{A}$ onto the eigenspace,

$$\boldsymbol{b}_{\mathcal{A}} = P_{\mathcal{A}}^{T}(\mathcal{A}(\boldsymbol{s}) - \bar{\mathcal{A}}). \tag{3.14}$$

$\boldsymbol{b}_{\mathcal{A}}$ only needs to be calculated once after the shape fitting has converged. The landmarks are predicted by implementing the LK algorithm in the orthogonal space [35],

$$\hat{\boldsymbol{s}} = \arg\min ||\mathcal{A}(\boldsymbol{s}) - \bar{\mathcal{A}}||^2_{\text{span}(P_{\mathcal{A}})} = \arg\min ||(\mathcal{A}(\boldsymbol{s}) - \bar{\mathcal{A}})^{\perp}||^2, \tag{3.15}$$

where $(\cdot)^{\perp}$ denotes the projection onto the orthogonal space, i.e., $(\cdot)^{\perp} = (I - P_{\mathcal{A}}P_{\mathcal{A}}^{T})(\cdot)$, with I being an identity matrix. In this way the salient appearance variations have been removed and a more robust LK method achieved. Equation (3.15) can be solved by iteratively linearising and inverse gradient descent by reversing the roles of the image and template [119],

$$\Delta\hat{\boldsymbol{s}} = \arg\min ||\bar{\mathcal{A}}^{\perp}(\Delta\boldsymbol{s}) - \mathcal{A}^{\perp}(\boldsymbol{s})||^2. \tag{3.16}$$

We apply the conditional independence assumption to simplify the calculation, i.e. the patches at the i-th landmark are only related to $\boldsymbol{x}_i$, therefore the equation can be decomposed into a set of independent equations,

$$\Delta\hat{\boldsymbol{x}}_{i,l} = \arg\min \left(\bar{A}^{\perp}_{i,l}(\Delta\boldsymbol{x}_i) - A^{\perp}_{i,l}(\boldsymbol{x}_i)\right). \quad i \in \{1, ...N\},\ l \in \ell_i \tag{3.17}$$

where $A_{i,l}$ is the feature patch at i-the landmark with l-th scale, flattened into a 1D vector. $\Delta\hat{\boldsymbol{x}}_{i,l}$ is the predicted increment of the i-th landmark inferred from $A_{i,l}$.

The solution is given by a least squares method,

$$\Delta\hat{\boldsymbol{x}}_{i,l} = \left(\frac{\partial \bar{A}_{i,l}^{\perp}}{\partial \boldsymbol{x}_i}\right)^{+} (A_{i,l}(\boldsymbol{x}_i) - \bar{A}_{i,l})^{\perp}, \tag{3.18}$$

in which $(\cdot)^{+}$ denotes the Moore-Penrose pseudo-inverse. Inside the bracket of the first factor is the gradient map of the mean patch at the i-th landmark and l-th scale, projected onto the orthogonal space.

Suppose we also have the variance $\boldsymbol{\sigma}_{i,l}^2$ of the prediction $\Delta\hat{\boldsymbol{x}}_{i,l}$, which could indicate the salience of the feature or the confidence of the prediction. To keep it simple, we calculate the variance as the mean squared difference between the patch observation and the template. Using a Gaussian parametric form and applying the product combination in (3.10), the likelihood of the location of the i-th landmark given the multi-scale predictions can be represented by,

$$p(\boldsymbol{x}_i|\{A_{i,l}\}_l) = \prod_l \mathcal{N}(\boldsymbol{x}_i; \hat{\boldsymbol{x}}_{i,l}, \boldsymbol{\sigma}_{i,l}^2), \tag{3.19}$$

where $\hat{\boldsymbol{x}}_{i,l}$ are the updated landmark estimated by adding $\Delta\hat{\boldsymbol{x}}_{i,l}$ to the current location. The advantages of combining the multi-scale predictions are given in section 3.2.1. We show next how to incorporate the predictions into a shape regulariser.

3.3.2 Shape regularisation

The shape can be either bounded by a subspace constraint [120] as in standard ASMs or optimised by a regulariser using, e.g., density estimation [64, 121], a Bayesian model [50], or sparse shape composition [122, 123], leading to more efficient fitting. It has been shown that utilising multiple predictions of individual landmarks can result in robust fitting. For example in [124], multiple candidates at a landmark are generated, then the best one is selected and the others are regarded as false positives. There have been multi-scale shape models [125, 126] to characterise the population

variations in a more accurate and robust way. To keep our method simple, we show how the single scale shape prior model introduce in section 3.1 can be applied in our multi-scale appearance model.

We assume that all of the multi-scale predictions from LFPs are valid, but with various weights across the landmarks and scales controlled by their variances, and deduce a regulariser to obtain the ML shape with respect to the shape prior and the multi-scale landmark predictions. Specifically, the likelihood of a shape instance given the shape prior Ω and image observation I can be represented as $p(\boldsymbol{s}|\Omega, I)$. Since Ω and I are conditionally independent, from Bayesian theory we have,

$$p(\boldsymbol{s}|\Omega, I) \propto p(\boldsymbol{s}|\Omega)p(\boldsymbol{s}|I). \tag{3.20}$$

The shape prior term is given by (3.9),

$$p(\boldsymbol{s}|\Omega) \propto \exp(\boldsymbol{b}^T \Lambda \boldsymbol{b}) \exp\left(\frac{1}{\rho}(||\boldsymbol{s} - \bar{\boldsymbol{s}}||^2 - |||\boldsymbol{b}||^2\right) \tag{3.21}$$

where $\Lambda = \text{diag}([\lambda_1, ..., \lambda_t])$. In a DAP we obtain shape observations from multiple scales by (3.19), and at each scale a subset of landmarks is estimated. In order to infer the optimal shape from this information, we first consider the two following questions:

1. At a certain level l, given the observation of a subset of landmarks in $\mathcal{K}_l$, how the whole shape can be deduced based on the observation and shape prior;

2. Given multiple predictions of a shape, how can the ML shape in terms of these predictions be calculated.

Inferring the whole shape from a subset of landmark estimations. At a single scale l of a trimmed DAP, we can only obtain the estimates of a subset of 'key' landmarks, $\hat{\boldsymbol{x}}_i$, $i \in \mathcal{K}_l$ with variances $\{\boldsymbol{\sigma}_i^2\}$. To estimate the whole shape from

this information, the remaining 'empty' landmarks can be inferred based on the key landmarks and the shape prior. Specifically, as we have no observation of the empty landmarks, their likelihood can be modelled as a Gaussian with infinite variance, which assumes all locations are equally likely. In this way we can write the likelihood for all landmarks observed from scale l as,

$$p(\boldsymbol{x}_i|I) = \begin{cases} \mathcal{N}(\hat{\boldsymbol{x}}_{i,l}, \sigma_{i,l}^2), & i \in \mathcal{K}_l \\ \mathcal{N}(\boldsymbol{0}, \text{Inf}) & i \notin \mathcal{K}_l. \end{cases} \tag{3.22}$$

Accordingly the shape observation becomes,

$$p(\boldsymbol{s}|I) = \prod_{i=1}^{N} p(\boldsymbol{x}_i|I). \tag{3.23}$$

Substituting (3.21) and (3.23) into (3.20) and taking the negative log form we can obtain an energy function,

$$E(\boldsymbol{s}) = \frac{1}{2}\boldsymbol{b}^T\Lambda^{-1}\boldsymbol{b} + \frac{1}{2\rho}(||\boldsymbol{s} - \bar{\boldsymbol{s}}||^2 - ||\boldsymbol{b}||^2) + \sum_{i=1}^{N} \frac{(\boldsymbol{x}_i - \hat{\boldsymbol{x}}_{i,l})^2}{2\boldsymbol{\sigma}_{i,l}^2}, \tag{3.24}$$

where $\hat{\boldsymbol{x}}_{i,l}$ takes the value zero and $\boldsymbol{\sigma}_{i,l}^2$ infinite at empty landmarks. The ML shape inferred from a single scale observation can be calculated by minimising this energy function. The resulting shape is the one best fitting the prior and the key landmarks. Fig. 3.8(a) gives an illustration of ML shape inference in the eigenspace.

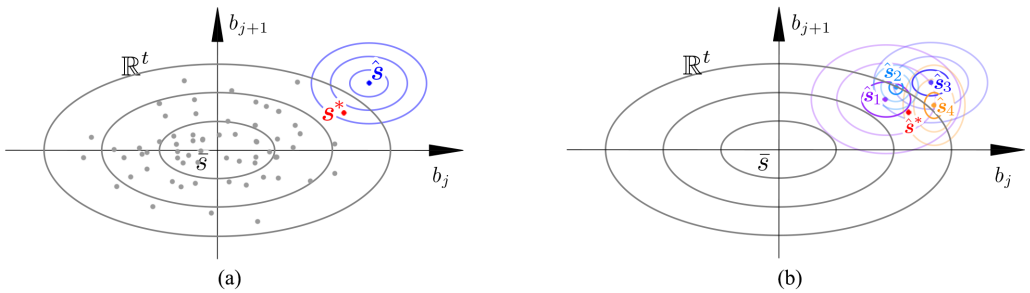


Figure 3.8: (a) An illustration of shape inference in the eigenspace spanned by $P \in \mathbb{R}^t$. Grey dots represent the training samples, ellipses show the three standard deviations of the Gaussian distribution which gives the prior knowledge of shapes. The shape observations $\hat{\boldsymbol{s}}$ are shown in blue, with the variance representing the confidence. The ML shape $\boldsymbol{s}^*$ is inferred from the prior and the observation. (b) ML shape inferred from the prior (grey) and multiple observations $\hat{\boldsymbol{s}}_i$. Ellipses show three standard deviations. The ML shape $\boldsymbol{s}^*$ is inferred seeking a balance between the prior and the observations.

Inferring the ML shape from multiple shape observations. Given multiple shape observations $\hat{\boldsymbol{s}}_i$ the likelihood of the shape can be formularised as a product,

$$p(\boldsymbol{s}|I) = \prod_{l=1}^{L} p_l(\boldsymbol{s}|I) = \prod_{l=1}^{L}\prod_{i=1}^{N} \mathcal{N}(\boldsymbol{x}_i; \hat{\boldsymbol{x}}_{i,l}, \boldsymbol{\sigma}_{i,l}^2). \tag{3.25}$$

Substituting (3.21) and (3.25) into (3.20) and taking the negative log form, a new energy function obtained is,

$$E(\boldsymbol{s}) = \frac{1}{2}\boldsymbol{b}^T\Lambda^{-1}\boldsymbol{b} + \frac{1}{2\rho}(||\boldsymbol{s} - \bar{\boldsymbol{s}}||^2 - ||\boldsymbol{b}||^2) + \sum_{l=1}^{L}\sum_{i=1}^{N} \frac{(\boldsymbol{x}_i - \hat{\boldsymbol{x}}_{i,l})^2}{2\boldsymbol{\sigma}_{i,l}^2}. \tag{3.26}$$

The shape minimising (3.26) is the ML shape with respect to the prior and all the landmark observations available. It seeks an optimal solution balanced between the prior and the observations, the weights of which are determined by the confidence of observations, see Fig. 3.8(b) for an illustration.

Instead of numerical optimisation, observing the homogeneous form of the

equation, we derive a closed-form solution,

$$\boldsymbol{s} = A^{-1}B + \bar{s}, \tag{3.27}$$

in which,

$$\begin{aligned} A =& P\Lambda^{-1}P^T + \frac{1}{\rho}(I - PP^T) + \sum_{l=1}^{L} \Sigma_l^{-1}, \\ B =& \sum_{l=1}^{L} \Sigma_l^{-1}(\hat{\boldsymbol{s}}_l - \bar{\boldsymbol{s}}), \end{aligned} \tag{3.28}$$

where $\Sigma_l = \text{diag}([\boldsymbol{\sigma}_{1,l}^2, ..., \boldsymbol{\sigma}_{N,l}^2])$. The proof can be found in Appendix B. The implementation of the explicit fitting algorithm is outlined in Algorithm 1.

Algorithm 1: Active Appearance Pyramid fitting

Training
1. Train the shape prior, obtain the mean shape $\bar{\boldsymbol{s}}$, the eigenvalues $\{\lambda_j\}_{j=1}^t$ and the eigenvector matrix P;
2. Build the Gaussian pyramid of training data and extract the training DAPs $\{\mathcal{A}\}$;
3. Train the appearance prior on $\{\mathcal{A}\}$, obtain the mean $\bar{\mathcal{A}}$ and the eigenvector matrix $P_{\mathcal{A}}$;
4. Calculate the mean in orthogonal space $\bar{\mathcal{A}}^{\perp}$ and the gradient of each patch in $\bar{\mathcal{A}}^{\perp}$, i.e., $\partial \bar{A}_{i,l}^{\perp}/\partial \boldsymbol{x}_i, i \in \{1, ...N\}, l \in \ell_i$.

Testing
1. Build the Gaussian pyramid of the testing image, initialise the shape $\boldsymbol{s}$;
2. Extract the DAP $\mathcal{A}(\boldsymbol{s})$ at the current shape;
3. **Local searching:** Project $\mathcal{A}(\boldsymbol{s})$ onto the orthogonal space, calculate the multi-scale landmark predictions by (3.18);
4. **Regularisation:** Calculate the ML shape $\boldsymbol{s}$ by (3.27);
5. Repeat 2 to 4 until the shape has converged;

Reconstruction of the object appearance. As the shape of the object is fitted using the method presented above and the appearance is encoded in the parameters $\boldsymbol{b}_{\mathcal{A}}$, we can recover the object information from the parameters. The reconstructed object can be visualised by first recovering the 'shape-free' appearance $\mathcal{A}$ by (3.11)

and then padding the multi-scale patches in $\mathcal{A}$ at the corresponding position, with the smaller scales layered on top of larger ones.

3.4 Experiments on 2D Lumbar Vertebral Images

To validate the performance by DAP we run experiments on 200 studies with dense landmarks and class labels. Cross-fold validation is performed. For assessing quantitative performance in landmark detection, we measure Point to Boundary Distance (PtoBD) and Dice Similarity Coefficients (DSC). For comparative analysis, we run the same data using implementations of AAM and CLM to assess convergence range, segmentation precision and, reconstruction appearance with AAMs. The performances of pathology classification on central canal stenosis and foraminal stenosis are also compared.

3.4.1 Experimental settings

Parametrisation. For axial images, the DAP is built with four level feature pyramids (see Fig. 3.6(b)). The patch size is 15×15 pixels. Similarly, for parasagittal images we use a three level DAP with the patch size of 9×9 pixels. In order to visualise the statistical variation among the population caused by LSS, we concatenate the appearance parameters $\boldsymbol{b}_{\mathcal{A}}$ in (3.14) and shape parameters $\boldsymbol{b}$ in (3.8) appropriately weighted for an equivalent variance. PCA is then applied to obtain the joint model. Fig. 3.9 shows the mean and the most significant variation of axial intervertebral anatomies L3/4 and L5/S1. Fig. 3.10 gives the mean and the first variation of the three intervertebral foramina. The first mode obtained by standard AAM reconstruction is also given in these cases for comparison. We can see that the DAP preserves more delicate features and richer information.

Validation. We run the two-fold cross-validation on each of the three L3/4, L4/5, L5/S1 axial datasets and three parasagittal dataset as introduce in Chapter 1. The

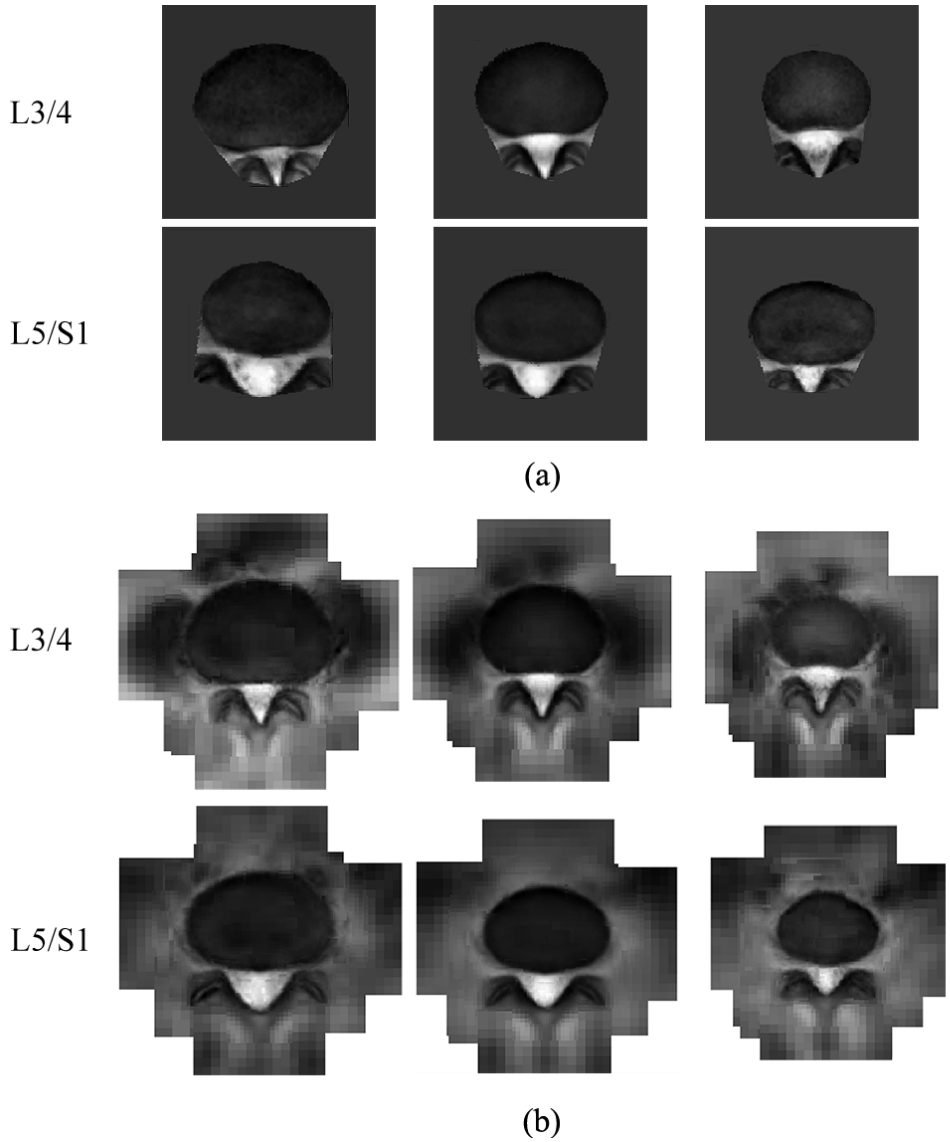


Figure 3.9: First mode of variation across the population with varied LSS, generated by (a) AAM and (b) DAP. The average appearance (middle) and the ± 2 SD variation are shown. Images are shown at the same scale. The DAP preserves more delicate texture of important features and covers a larger contextual region.

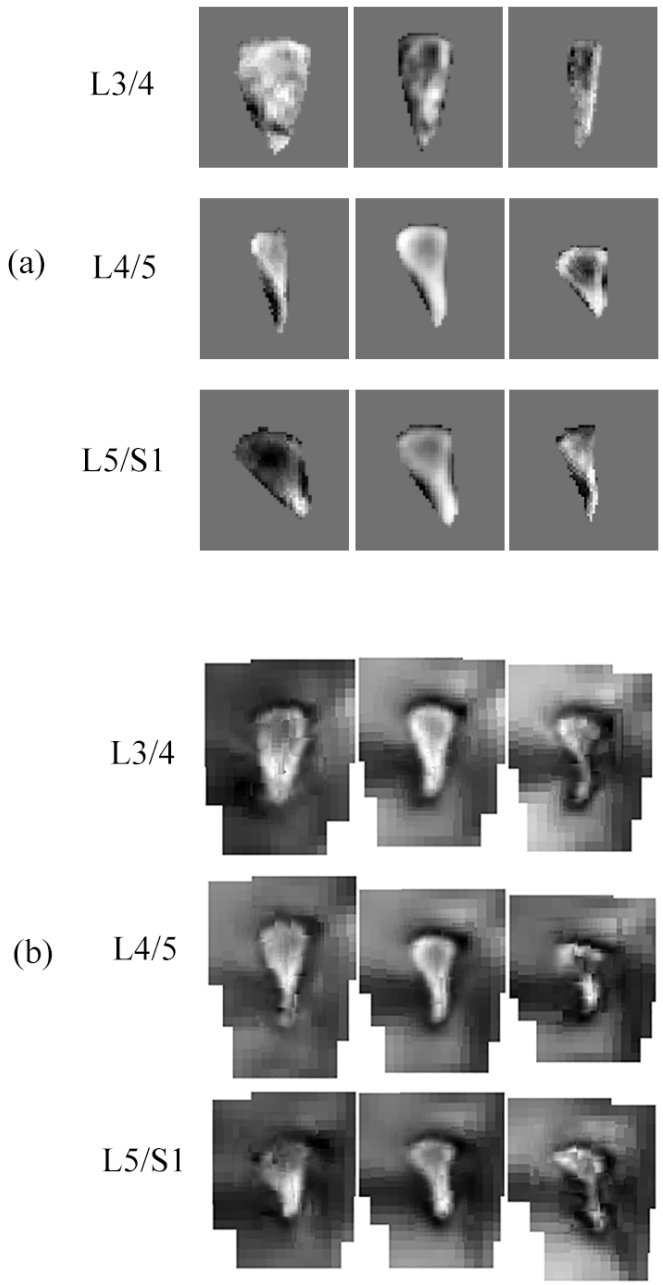
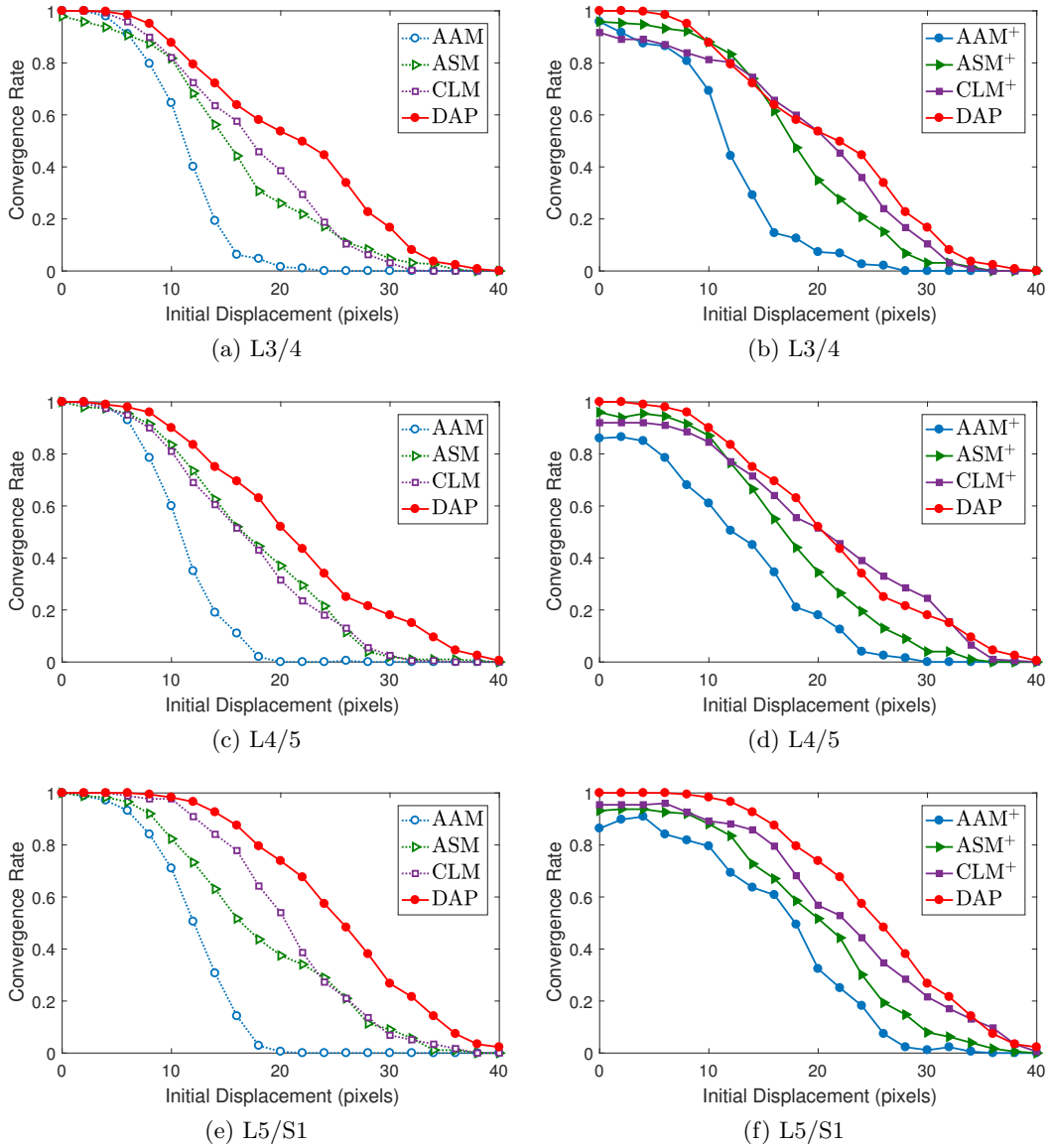


Figure 3.10: First mode of the variation of the three foramina generated by (a) AAM and (b) DAP. The mean (middle) and the ±2 SD variation are shown. Images are shown at the same scale.

validation is repeated several times to obtain unbiased results. Two measurement criteria are used for the evaluation: the PtoBD in pixels and the DDSC [127]. DSC is defined as the amount of the intersection between a segmented object and the ground truth, DSC = $2 \cdot \mathrm{TP}/(2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN})$, with TP, FP, FN denoting the true positive, false positive and false negative values respectively. For the axial images, the DSC of the canal and disc contours between the fitted shape and the ground truth is used as the criterion of segmentation precision. We compare the proposed DAP with three popular methods: AAMs [41] as a standard holistic method, ASMs as a widely used shape model, and CLMs [50] as a popular part-based approach. For consistency, in the CLMs we use the same patch size as in the DAP. The AAM model based on the inverse compositional image alignment algorithm proposed in [41] is a modified version of the conventional AAM models [10]. The fitting accuracy is reported as a baseline. CLMs have shown impressive performance in object detection as well as in medical image segmentation. The implementation presented in [50] is chosen for comparison as a promising performance has been demonstrated on medical images.

3.4.2 Results

Convergence range. We run displacement experiments on the axial images to test the convergence performance of the three methods. The shape of each testing image is initialised as the mean shape with displacement from the true location in four directions. The searching algorithms are then applied to the image. We say a case converges if the final DSC is larger than a threshold value, which is set to be 0.8 as an example. Fig. 3.11 shows the proportion of converged cases with different initial displacements on L3/4, L4/5 and L5/S1 respectively. The compared methods are AAM, ASM and CLM as well as their coarse-to-fine implementations at three scales. We can see in Fig. 3.11(a)(c)(e) that DAPs have a significantly larger convergence range over all three methods. In Fig. 3.11(b)(d)(f) we observe that although coarse-to-fine implementations can improve the overall convergence range



(a) L3/4 (b) L3/4

(c) L4/5 (d) L4/5

(e) L5/S1 (f) L5/S1

Figure 3.11: Successful convergence rate of compared methods on lumbar intervertebral slices L3/4, L4/5 and L5/S1. LEFT: comparison with the single scale methods. RIGHT: comparison with the coarse-to-fine version of these methods (denoted by $(\cdot)^+$). DAP shows a superior performance in convergence range against all three methods, as well as robustness against the coarse-to-fine implementation of these methods.

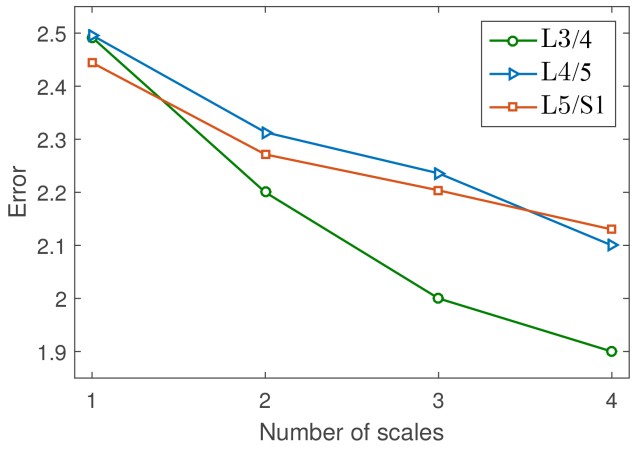
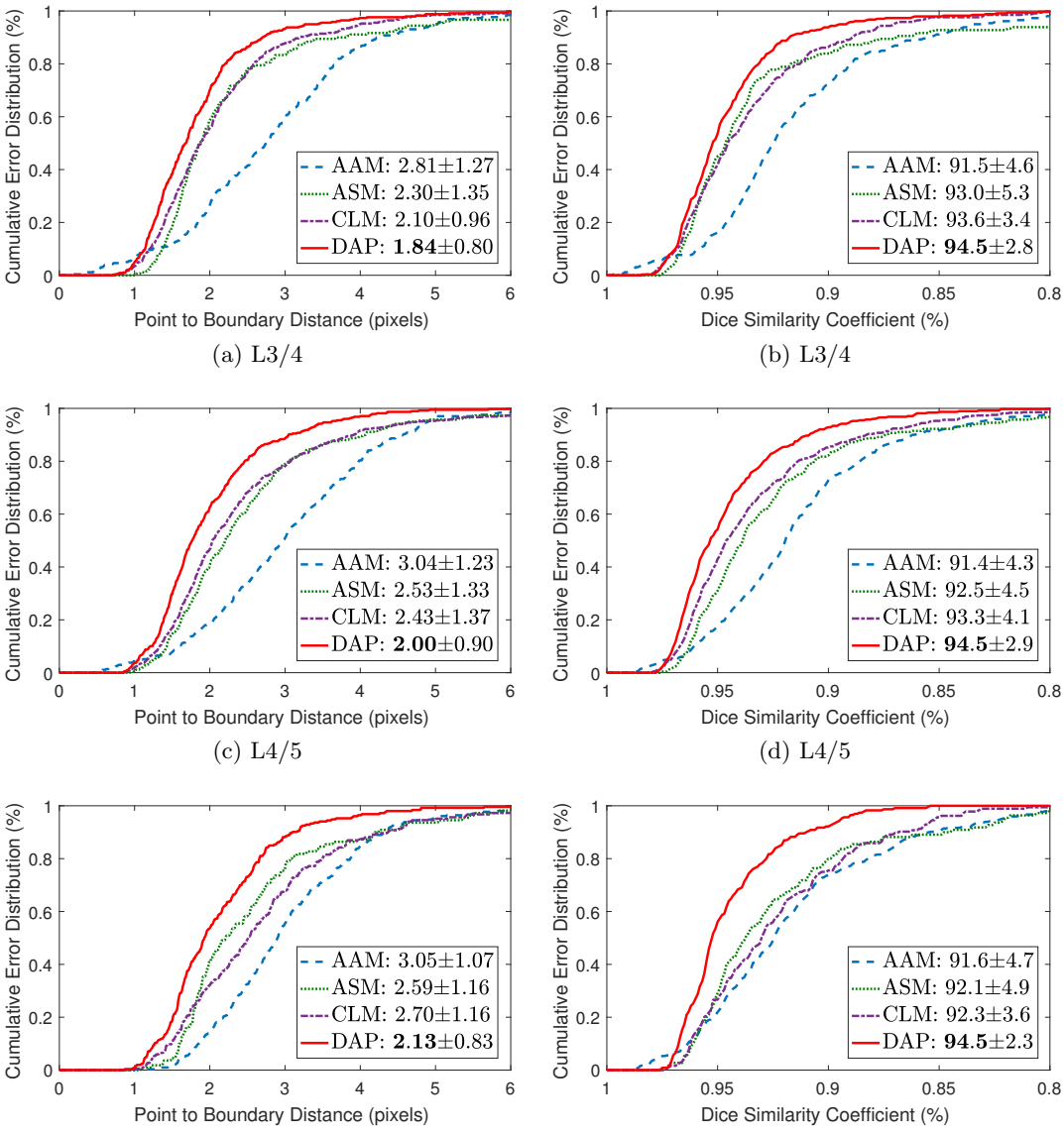


Figure 3.12: Fitting error against the number of scales used in DAP.

of the three methods, the failure rate increases as well. For example they have much lower successful convergence rates at the zero initial displacement, which means in low quality or challenging cases, the shape could diverge at the coarse level because of lack of texture details. This further supports our argument of combining multi-scale features to enhance the robustness. The improvement of DAP is on account of the multi-scale LFPs: the larger scales ensure a wider capture range, while the smaller scales take effect as soon as it gets into the convergence range.

Accuracy of segmentation. For each testing case, the shape is initialised as the mean shape with a three-pixel displacement from the true position in random directions. To demonstrate the benefit of using the multi-scale local feature pyramids as feature descriptor, we report the performance of DAP with different number of scales in Fig. 3.12. We can see that in all three subsets the fitting error reduces with the increasing number of scales utilised.

Due to the higher failure rate of the coarse-to-fine approaches even at small initial displacements (as shown in Fig. 3.11), we only compare the accuracy of our DAP with the single scale implementation of these methods, and set the initial displacement small enough to keep them within a confident convergence range. The algorithms are then applied to fit the shape to the image. We repeat the process



(a) L3/4 (b) L3/4

(c) L4/5 (d) L4/5

(e) L5/S1 (f) L5/S1

Figure 3.13: Cumulative error distributions of segmentation of lumbar intervertebral slices: L3/4, L4/5 and L5/S1. DSC and DPtoB (in pixels) are used as the criteria. Compared methods are AAM, ASM, CLM and DAP. The legends give the mean errors and standard deviations.

several times for an unbiased result. The cumulative error distribution of the DSC and PtoBD of the segmentation results on three axial dataset are shown in Fig. 3.13. The mean error and one standard deviation (SD) is also given in the legends for the comparison. We can see that DAP achieves the best precision of segmentation. Meanwhile the smaller SD shows that DAP has the superior consistent performance, which is also indicated in the cumulative error distribution curves.

The qualitative results of segmentation on five representative cases are shown in Fig. 3.14, with the difficulty increasing from left to right. The ground truth shape is shown in each case for convenience. We can see that the AAMs, ASMs and CLMs are affected by local ambiguity (highlighted by red circles) on the challenging cases and become trapped in a local minimum. We observe a large proportion of outliers by AAM around the disc like the third case in Fig. 3.15. A possible reason is that the plain textures inside the disc contain very limited information. The DAPs shows a robust and consistent performance in all five cases.

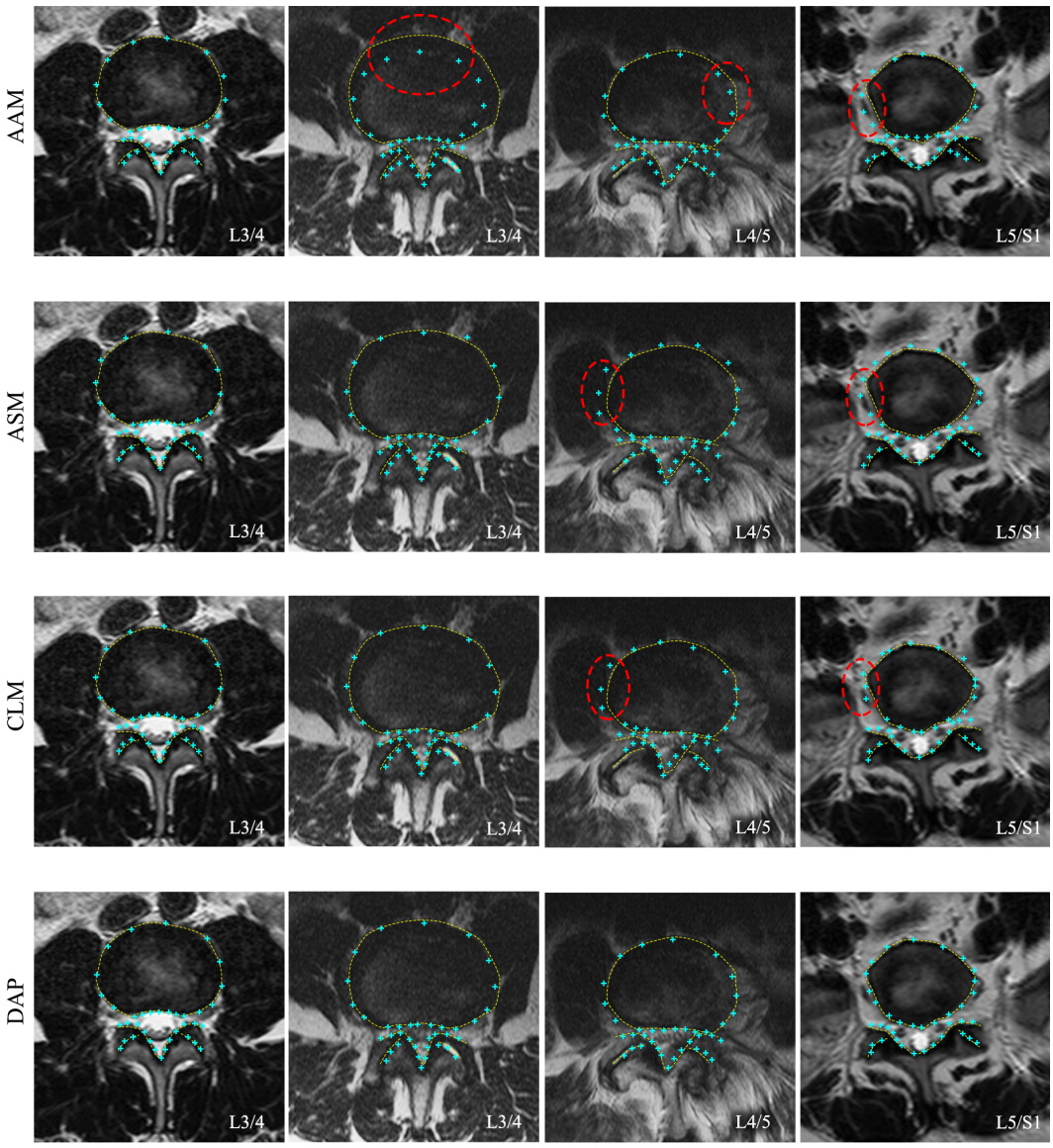


Figure 3.14: Segmentation results on four cases, increasing in difficulty from left to right. The ground truth of segmentation is shown by yellow dash lines, fitting results are shown by cyan crosses. Red circles highlight the outliers.

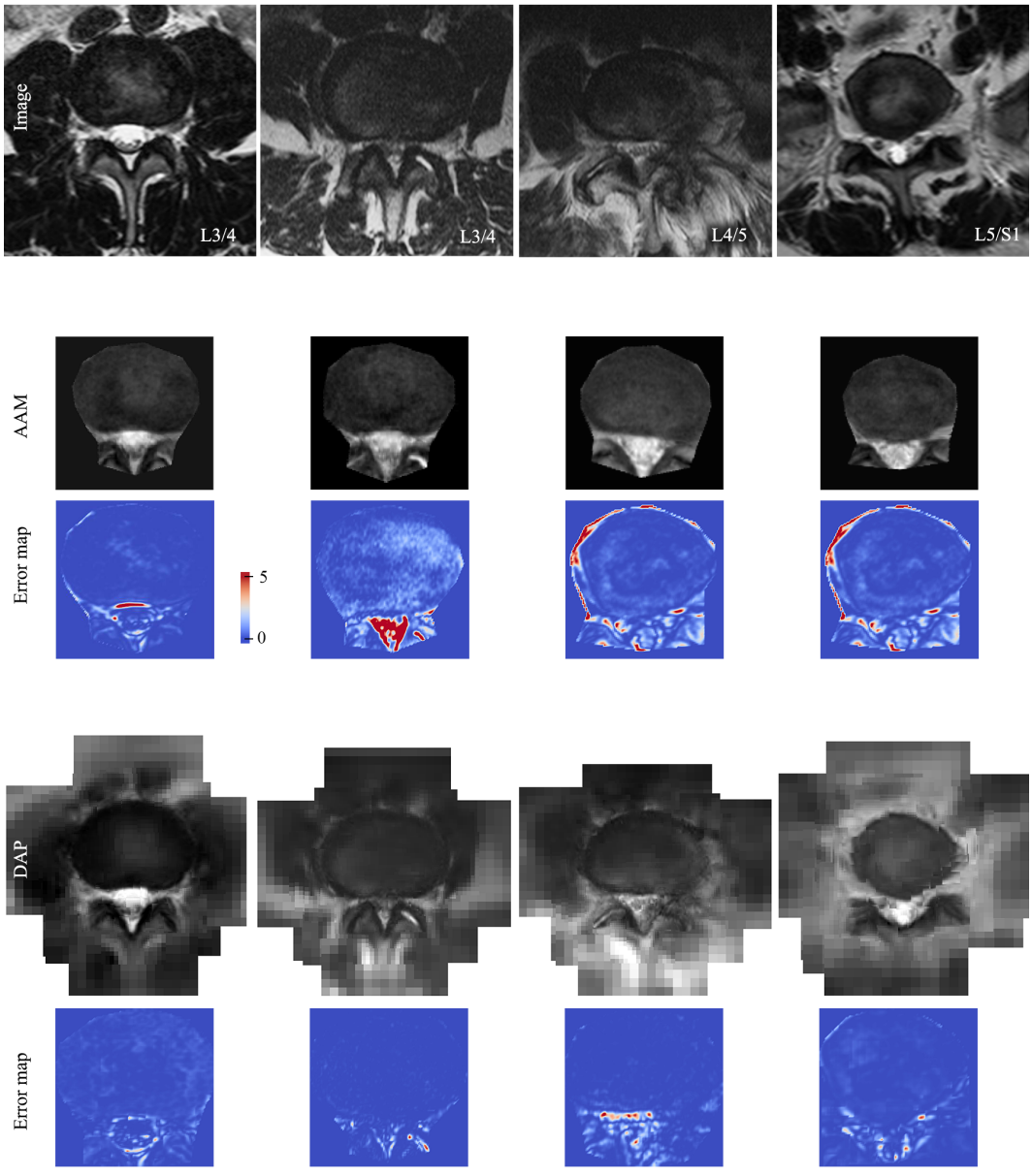


Figure 3.15: Reconstruction results on four cases, increasing in difficulty from left to right. The reconstructed appearances are the parametric models with the parameters fitted to the instances. The error maps highlight the regions with low fitting precision, which are mainly around the features of interest.

Comparisons of object reconstruction. As the parameters of AAM and DAP encode both the shape and appearance information, we can reconstruct the anatomy from the fitted parameters. In addition to morphometric comparison, the quality of appearance synthesis can indicate how precise the object is modelled and appearance details are represented. We therefore quantify and compare the appearance fitting quality using image distortion as a measurement. We calculate the error map of a synthesised appearance as follow,

$$\mathrm{Err}(\boldsymbol{x}) = \frac{[I(\boldsymbol{x}) - J(\boldsymbol{x})]^2}{[I(\boldsymbol{x})]^2}. \tag{3.29}$$

where I is the true image and J is the synthesised result. The synthesised appearance as well as the error map for five cases by AAM and DAP are shown in Fig. 3.15.

We can see that DAP preserves finer structural details and covers larger area of contextual information. For example, the facet is precisely located and the facet texture is well preserved in all five cases. In case three and four, the DAP delineates the degenerated vertebrae and the compressed central canal more accurately than AAM does. The large errors of AAM are mainly distributed around the feature of interest where the pathology might appear. We also evaluate the overall synthesis error of a case by calculating the signal-to-noise ratio (SNR),

$$\mathrm{SNR} = \frac{E[I(\boldsymbol{x})]^2}{E[I(\boldsymbol{x}) - J(\boldsymbol{x})]^2}, \quad \boldsymbol{x} \in \Omega, \tag{3.30}$$

where Ω is the region within the shape mesh as it is the region modelled by AAMs. The means and SD of the SNR of the testing samples are reported in Table. 3.1. We can see that compared with the shape fitting results, the improvement in appearance fitting by DAP is more significant.

Reconstruction of neural foreman. We also report the qualitative results of reconstruction of neural foramina on parasagittal images in Fig. 3.16. We observe

Table 3.1: Means and SD of SNR of synthesised results by AAM and DAP.

	L3/4	L4/5	L5/S1
AAM	4.80±2.73	5.36±2.60	7.51±5.06
DAP	**8.72**±4.71	**6.96**±3.77	**9.38**±4.71

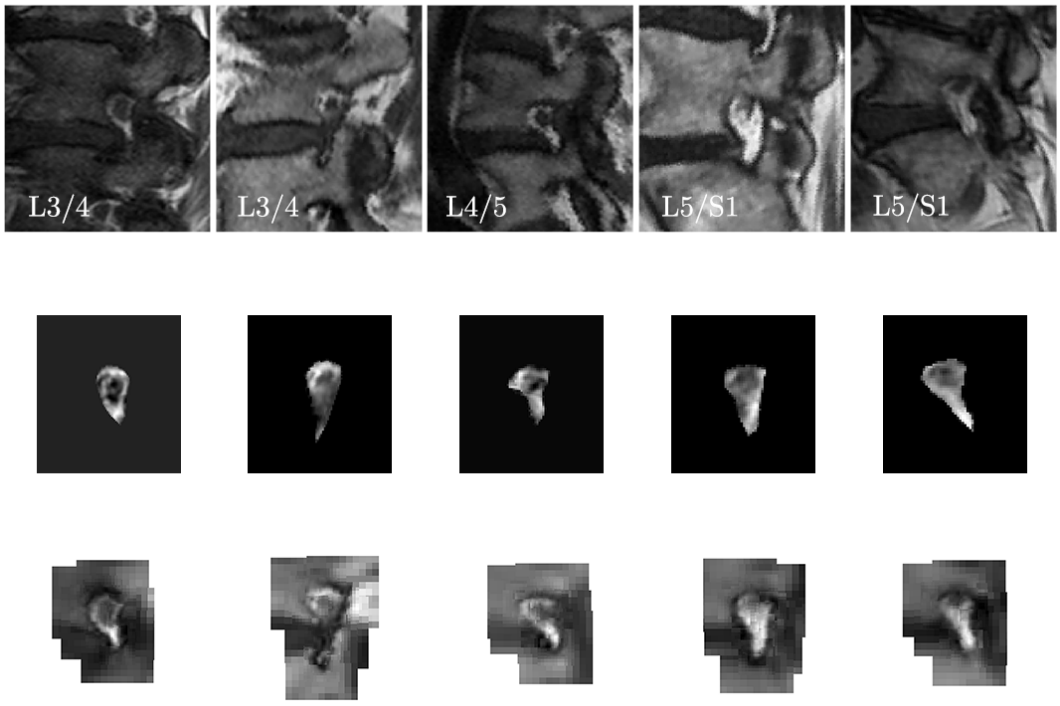


Figure 3.16: Fitting and reconstruction results of neural foramen on five cases. Top: Testing data; Middle: Reconstructed appearance by AAMs; Bottom: Reconstructed appearance by DAPs.

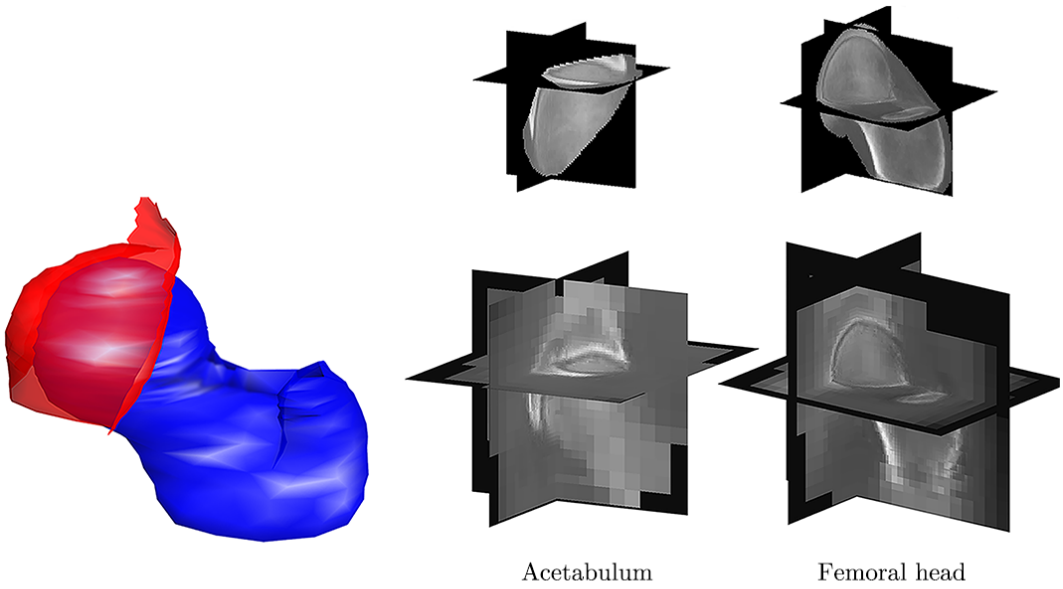
that the inner region of the foramen can provide very limited information for a robust fitting as they are nearly convex contours, which is the cause of the degraded performance of AAM.

3.5 Experiments On 3D Hip Joint Data

Data. To demonstrate the performance on 3D data, we build DAP models on CT volumes of the hip joints of 38 patients suffering from degrees of femoroacetabular impingement. We comparatively assess the computational cost against AAM, the mean surface errors and the reconstruction quality. The data are pre-interpolated to obtain an isotropic voxel size of 1 mm. The femoral head and acetabulum are annotated by 427 and 254 points marked up by experts. We build two DAP models delineating these two anatomies respectively. Both models are composed of four-level cubic patches with a consistent size of $9 \times 9 \times 9$ voxels. A cross validation is performed on 38 CT volumes, i.e., randomly picking 19 samples as training data, and testing on the remainder, and repeating.

Computational efficiency. The DAP model parametrising the femoral head consists of 617 patches with size of nine-voxels cubed, which is 449,793 voxels for each instance. As a comparison, the AAM uses a $92 \times 96 \times 96$ volume which consists of 847,872 voxels. Thus the DAP uses 53% of voxels compared with the AAM, while covering a much larger contextual region and preserving a full resolution of the features of interest such as the articular surface. Similarly, a second acetabulum model uses 58% of the voxels the AAM does.

We tested the time consumed by the AAM and DAP for training and fitting using a quad-core 3.2GHz processor with 16GB memory. Both algorithms were implemented in MATLAB, with the intensive computations of the AAM compiled in C++ language to boost its performance. We observe that it takes 170 ms to generate a shape-free appearance of femoral head by warping the volume, after compilation in C++. As a comparison, the most intensive computation of DAP, i.e., to generate the Appearance Pyramid by extracting subvolumes from the data, takes only 40 ms in MATLAB. We report the time consumed by each principal task



(a) (b)

Figure 3.17: (a) The mean shape of the acetabulum (red) and femoral head (blue). (b) The mean appearance of the two anatomies generated by AAM (top) and DAP (bottom).

on the femoral head data in Table 3.2. We can see that the DAP consumes less than 10% the training time and 15% the testing time of the AAM.

Table 3.2: Time consumption of AAM and DAP on femoral head

Process	AAM		DAP
Loading data:	18.0 s		18.0 s
Training:	9.8 min	Build gaussian pyramids:	45.8 s
		DAP training:	7.4 s
		Total:	**53.2** s
Fitting (30 iterations):	18.3 s		**2.7** s
Reconstruction:	0.6 s		**0.3** s

Precision of segmentation and reconstruction. We compare the performance of DAP with AAM in segmenting the femoral head and acetabulum. The mean shape of the two anatomies is shown in Fig. 3.17(a). The mean appearances generated by the AAM and DAP are given in Fig. 3.17(b). We calculated the vertex-to-surface

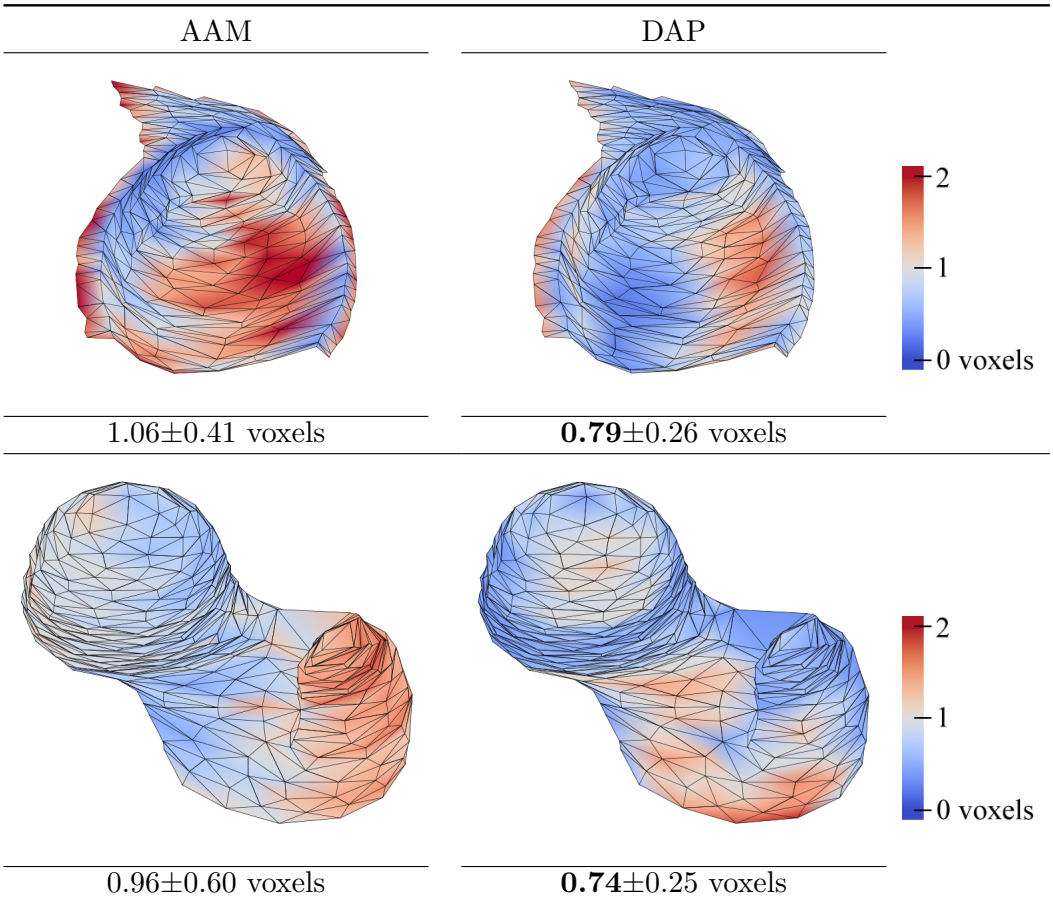


Figure 3.18: Mean vertex-to-surface errors of the segmentation results of the acetabulum and femur head, displayed on the mean shape mesh. The mean errors and standard deviations are shown at the bottom.

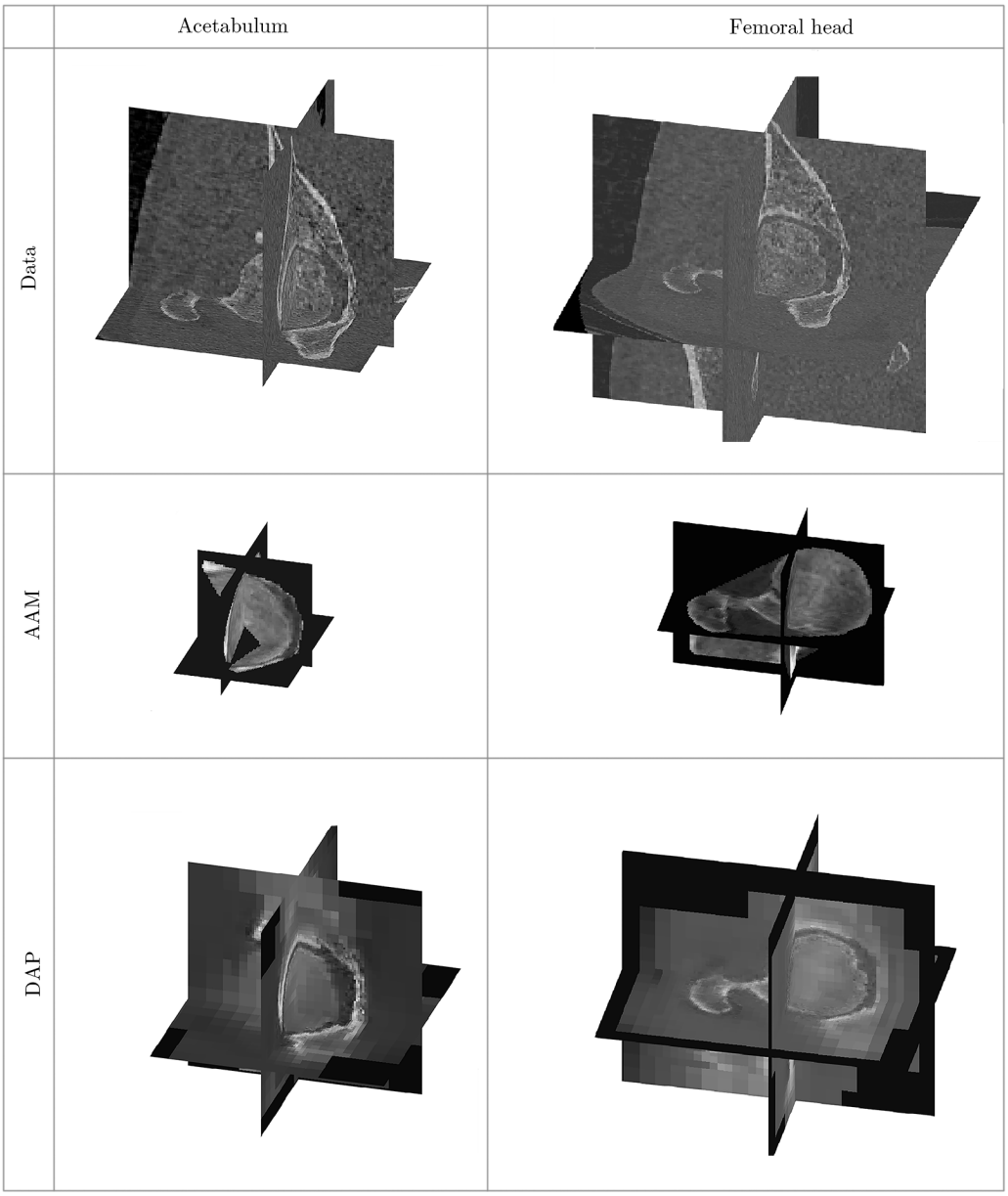


Figure 3.19: Qualitative results of the reconstruction. Shown are the testing data (top), and the appearance modelled and fitted by AAM (middle) and DAP (bottom). All images are shown at a same scale.

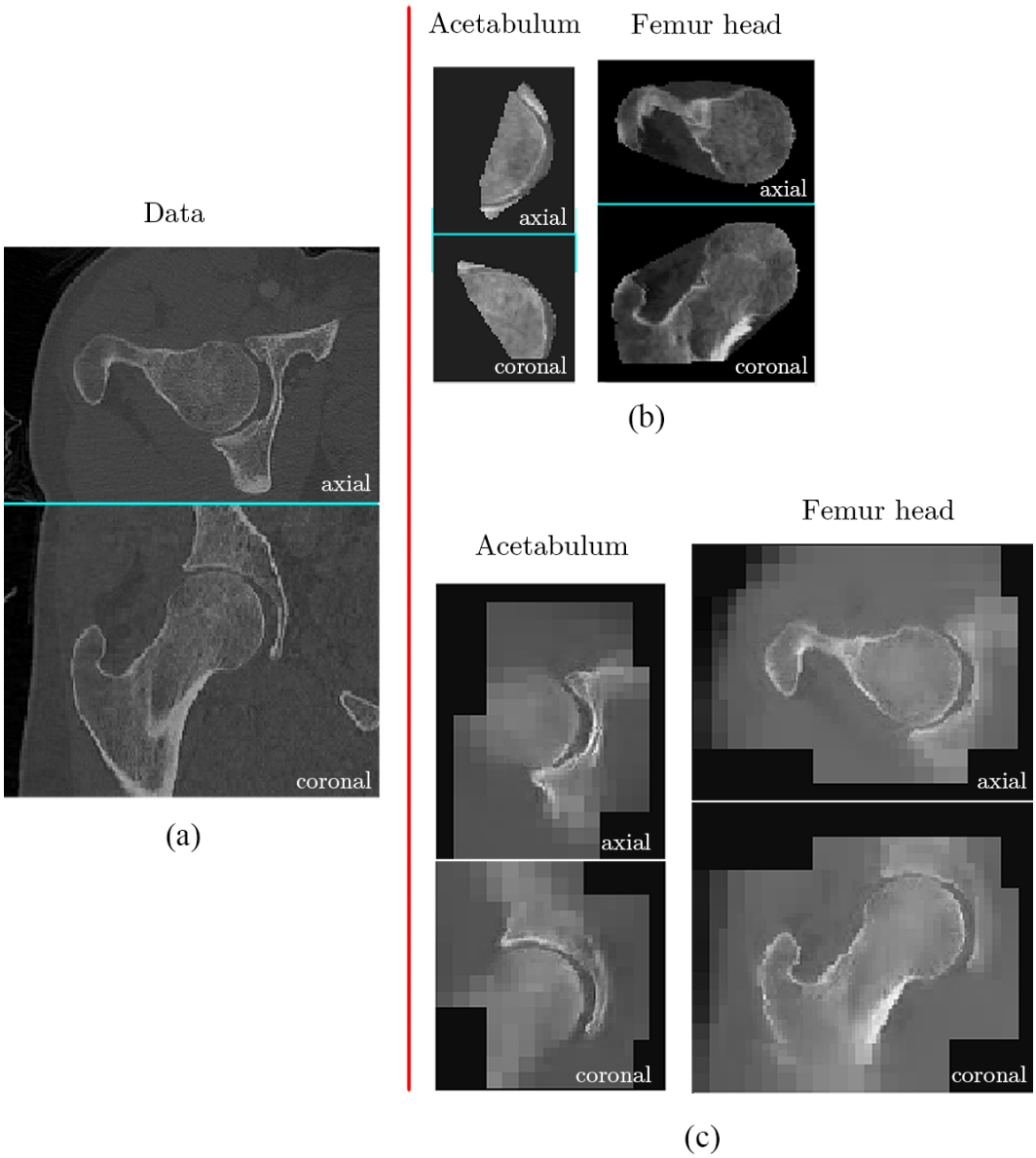


Figure 3.20: Qualitative results of the reconstruction. Shown are the testing data (a), and the appearance modelled and fitted by (b) AAM and (c) DAP. The volumes are shown with paired axial and coronal cross-sections. (Videos and DICOM files of the 3D results are available online at `https://sites.google.com/site/activeappearancepyramids/`.)

errors to assess the quantitative performance of the segmentation. The mean errors at individual vertices are visualised on the mean shape mesh in Fig. 3.18. The mean value of the overall errors and the SD across data and tests are also given at the bottom. We can see that the DAP has a significant smaller mean error: 0.79 voxels versus 1.06 voxels on the acetabulum, and 0.74 voxels versus 0.96 voxels on the femoral head. In addition, the smaller SD indicates the robustness of DAP across the cases. Note that the large fitting error by AAM is mainly distributed at the greater trochanter due to the irregular anatomical shapes. The DAP shows large fitting error at the bottom of the image volumes as no enough textural information can be acquired at the edge of the volumes.

Fig. 3.19 shows the fitting and reconstruction results of the acetabulum and femoral head on a case by the AAM and DAP respectively. Another case is shown with cross-sections in Fig. 3.20 to give a clearer view. Anatomies in each figure are shown in the same size ratio. The DAP syntheses cover a larger contextual region, which is why they appear to be larger. We can see that the DAP preserves sharper and more precise structures. Whereas in the AAM the reconstruction is blurred and with noticeable distortion.

3.6 Discussion and Conclusions

In this chapter we presented a part-based appearance model we refer to as an DAP. A simultaneous landmarks searching and appearance fitting algorithm was derived based on the weighted LK method. We introduced a shape regulariser utilising multi-level landmark estimation, and derive a closed-form solution to the maximum likelihood shape. The DAP can parametrise an object class and synthesise new instances as an AAM does. However the DAP differs from holistic AAMs in two respects:

1. AAMs model intra-class variations with local affine transforms, while DAPs

approximate the deformation with local translations of multi-scale parts;

2. AAMs model the inner region of the shape mesh while DAPs cover the contextual information with multiple resolutions.

We ran experiments to validate its performance and highlighted its advantages in several respects:

1. Computational efficiency. Computational cost has been a main limitation in existing appearance models tackling volume data. Compared with the AAMs, an DAP keeps full resolution of salient features, with reducing resolution further away from landmarks, which covers larger context but consumes less memory. DAP training and fitting is much faster because no image warping or interpolation is needed. The time consumption for both training and testing is linear to the number of samples in the dataset, so we would expect a time saving of 90% / 85% in training / testing correspondingly on large scale clinical data.

2. Fitting accuracy and robustness. The DAP spreads outside the shape mesh and captures more contextual information. Compared with AAMs and CLMs, the multi-scale feature descriptors enhance both position specificity and textural distinguishing ability, result in a superior fitting precision and robustness to local minima. In clinical practice we expect it to be able to deal with challenging cases better than conventional methods with lower failure rate. The larger convergence range also makes it less sensitive to initialisation, e.g., inputs of initial positions by clinicians.

3. Precision of parametrisation and reconstruction. We observe a finer and preciser reconstruction result in DAP. The better quality of reconstruction indicates two facts. Firstly, it captures and utilises more precise object appearance for shape fitting, which is demonstrated by its better segmentation performance. Secondly, it indicates that the more delicate and richer appearance

is parametrised and encoded in the DAP parameters. In addition, the better delineation results are beneficial for revealing the anatomical changes caused by either the diversity in population or the pathological degeneration, which can be useful for teaching and demonstration purposes.

The DAP presented in this chapter represents the appearance of objects with part models built on Gaussian image pyramids. In the next chapter, we show that more advanced image pyramids based on wavelet representations can be utilised to further improve the performance.

CHAPTER 4

Wavelet Appearance Pyramids

for Landmark Detection and Pathology Classification

Following on from the proposed DAP appearance model, in this chapter we propose to further decompose the parts in a DAP into multi-scale basic feature components, which is accomplished by replacing the Gaussian pyramids by advanced wavelet pyramids. We refer to the new appearance model as a Wavelet Appearance Pyramid [128]. To achieve an explicit scale decomposition, the filter banks are designed and arranged directly in the Fourier domain. The logarithmic wavelets (loglets) [30] are adopted as the basis functions of the filter banks for their superior properties, such as uniform coverage of the spectrum (losslessness) and infinite number of vanishing moments (smoothness). The scales are complementary in the Fourier domain which enables the reconstruction of the appearance from a WAP. The variations in the population can be modelled and visualised, with the deformation approximated by local rigid translations of the multi-scale parts, and the appearance changes by linear modes of the assembly of the parts. We introduce the composition and fitting

of WAP in details as follows. [1]

4.1 Object Representation with WAP

To provide a more comprehensive description of an object, we decompose the appearance into pyramidal channels at *complementary* scale ranges with wavelets, and represent each channel with a part-based model. The method is demonstrated in Fig. 4.1 and detailed as follows.

4.1.1 Explicit scale selection in the Fourier domain

We start by decomposing an image I into multi-scale channels directly in the Fourier domain. When considered in polar coordinates, the Fourier spectrum $\mathcal{I}$ actually spans a scale space with larger scales at lower frequency and smaller scales spreading outwards. Therefore a multi-scale decomposition of image textures can be achieved explicitly by dividing the spectrum into subbands, see Fig. 4.2(b). In practice, filtering the spectrum with sharp windows will introduce discontinuities therefore causing aliasing [2]. To design a bank of window functions which are smooth in shape while uniformly covering the spectrum, we use loglets as the basis functions because they possess a number of useful properties [30].

Denoting the frequency vector by $\boldsymbol{u}$ and its length by ρ, a bandpass window with a loglets basis can be designed in the Fourier domain as,

$$\mathcal{W}(\boldsymbol{u};s) = \mathrm{erf}\left(\alpha \log\left(\beta^{s+\frac{1}{2}} \frac{\rho}{\rho_0}\right)\right) - \mathrm{erf}\left(\alpha \log\left(\beta^{s-\frac{1}{2}} \frac{\rho}{\rho_0}\right)\right) \tag{4.1}$$

where α controls the radial bandwidth, s is an integer defining the scale of the filter, and $\beta > 1$ sets the relative ratio of adjacent scales – set to two for one octave intervals. ρ_0 is the peak radial frequency of the window with scale $s = 0$.

[1] Supplementary videos of this chapter can be found at `https://sites.google.com/site/appearancepyramids1`.

[2] See an example in Appendix C.

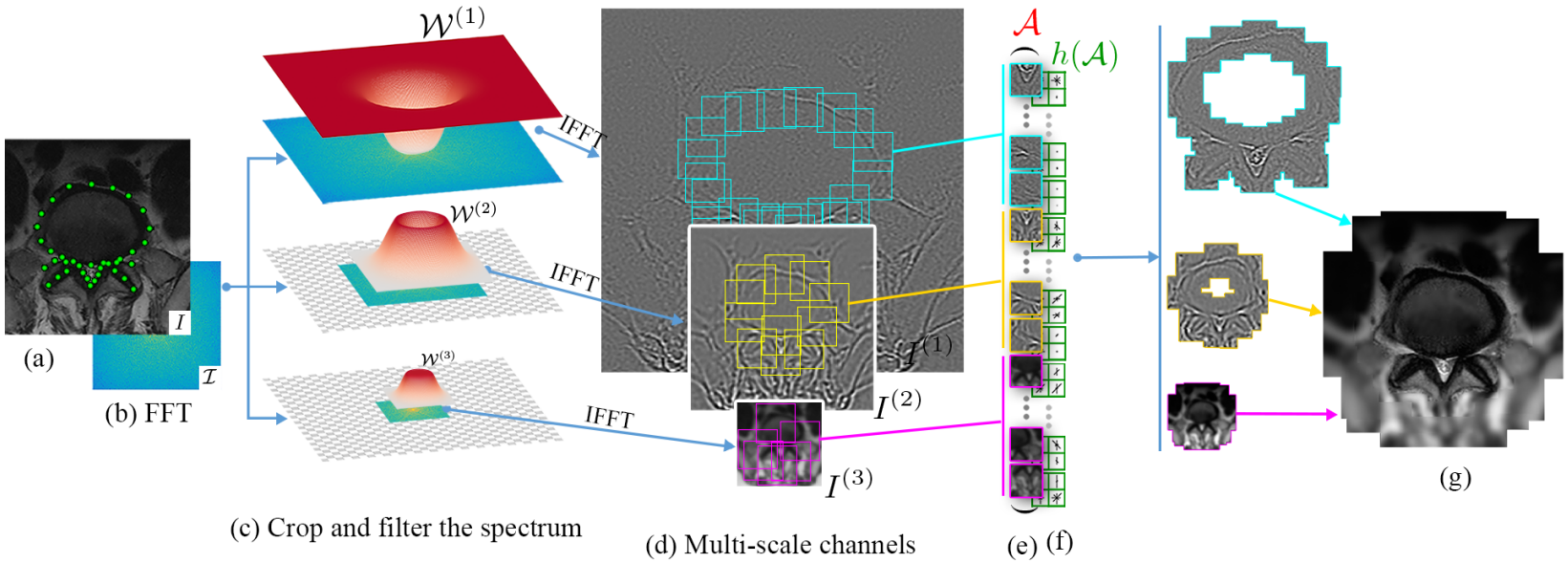


Figure 4.1: The proposed appearance model ($\mathcal{A}$) and feature descriptor ($h(\mathcal{A})$). **Decomposition:** (a) An image example. (b) Fourier transform of the image. (c) Multi-scale windows $W_x^{(l)}$ are applied to the spectrum. As the windows cover only a subband at one octave lower, spectrums are cut by half at each larger scales. (d) Subband pyramids representing multi-scale structures are obtained directly from the filtered spectrum, with a simultaneous downsampling at larger scale achieved by the cropping in the Fourier domain. (e) Local patches are extracted from the subband channels at key landmarks in $\boldsymbol{s}$. Patches at different channels have the save size in pixels, which give multi-scale description of the local features. (e) All patches are concatenated and flattened into 1D vector $\mathcal{A}$ serving as the profile of the appearance. **Reconstruction:** (f) Feature patches are padded at each scale level with spatial configurations $\boldsymbol{s}$. (g) All scales are accumulated to recover the object appearance. Note in (c) that the lowest frequency in the centre is not covered, therefore in (g) the background is removed. (h) The lowband channelcan be added to compare with the original image.

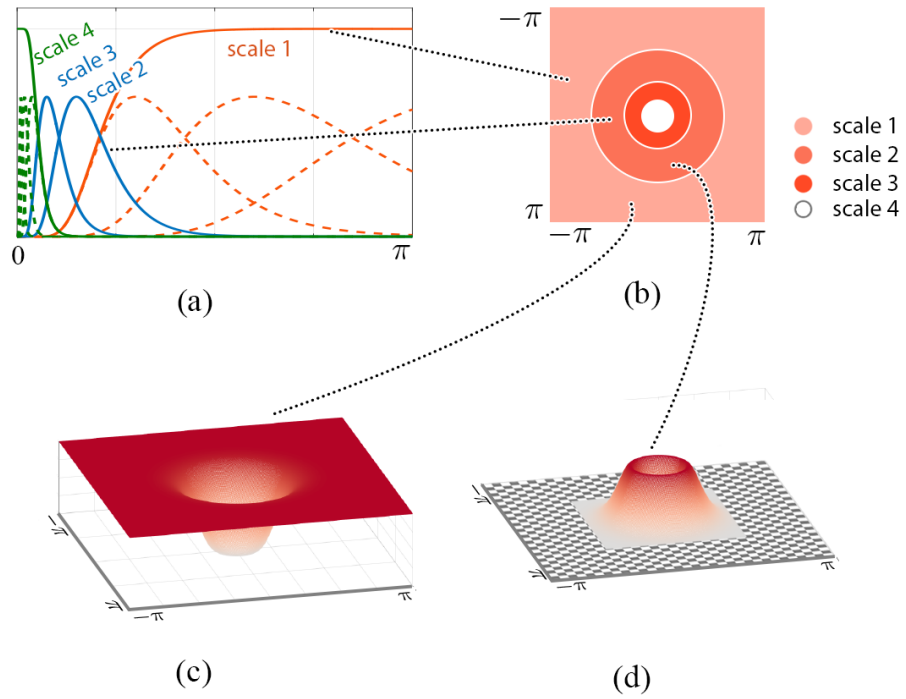


Figure 4.2: (a) Radial profiles of the filters. (b) Scale selection in the Fourier domain. (c) The high pass filter in the Fourier domain. (d) The first bandpass filter.

To extract the sharp textures of an image, the first scale channel should cover the highest frequency components. Noting the uniform property of the loglets, we accumulate a group of loglets successively having one-octave higher central frequencies as the first scale window, i.e., $\mathcal{W}^{(1)} = \sum_s \mathcal{W}(\boldsymbol{u}; s), s = \{0, -1, ...\}$, which achieves an even coverage towards the highest frequency, see the 1D profile in Fig. 4.2(a) shown as a red curve, and the 2D window in Fig. 4.2(c). The second and larger scale features can be selected by windows covering lower frequencies, $\mathcal{W}^{(s)}(\boldsymbol{u}) = \mathcal{W}(\boldsymbol{u}; s-1)$. Profiles of two adjacent larger scale windows are shown in Fig. 4.2(a) as blue curves, and a 2D window shown in Fig. 4.2(d). For a lossless decomposition, the largest scale window should uniformly cover the lowest frequencies, so it is designed as a accumulation of the remaining loglets functions, $\mathcal{W}^{(L)} = \sum_s \mathcal{W}(\boldsymbol{u}; s), s = \{L-1, L, ...\}$, see the green curve in Fig. 4.2(a). L is the total number of scales in the filter banks.

As the image filtering can be implemented in the Fourier domain by mul-

tiplication, the filters can be efficiently applied by windowing them on the image spectrum $\mathcal{I}$, and the image channels obtained by the inverse Fourier transform of the windowed spectrum, $I^{(s)} = \mathcal{F}^{-1}(\mathcal{I} \cdot \mathcal{W}^{(s)}), s = \{1, 2, ..., L\}$. The image is thus decomposed into complementary channels $\{I^{(s)}\}$.

4.1.2 Wavelet Appearance Pyramid

Wavelet image pyramid. It is evident that larger scale textures can be described sufficiently at a lower resolution. Note in Fig. 4.2(a) that the magnitude of the two larger scale windows beyond $\pi/2$ and $\pi/4$ is almost zero. Therefore we can discard these areas of the spectrum, which results in an efficient downsampling without information loss or aliasing effect [3]. As a result, the resolution is reduced by 2^s at scale s and a subband pyramid is obtained, see Fig. 4.1(c)(d).

Wavelet appearance pyramid. Given a landmark $\boldsymbol{x}$, we extract an image patch A_s at each scale s of the pyramid. All patches $\{A_s\}_{s=1}^{L}$ have the same size in pixels, which describe the local features at octave larger scales, domain sizes and lower resolutions, see Fig. 4.1(d)(e). A WAP, denoted by $\Phi = [\mathcal{A}, \boldsymbol{s}]$, consists of an assembly of feature patches $\mathcal{A} = \{\{A_{s,i}\}_{s=1}^{L}\}_{i=1}^{N}$ extracted at all the landmarks $\{\boldsymbol{x}_i\}_1^N$, and a shape $\boldsymbol{s} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ...\boldsymbol{x}_N]$ designating the locations of the patches. At larger scales fewer patches are manually chosen at key landmarks to reduce the overlapping. Φ is then flattened into a 1D vector serving as the profile of the anatomy. A further feature extraction function such as Histogram of Oriented Gradient (HOG) can be readily applied on the patches to reduce the dimensionality and enhance its robustness, i.e., $h(\mathcal{A}) = \{\{h(A_{s,i})\}_{s=1}^{L}\}_{i=1}^{N}$, see Fig. 4.1(f). To reconstruct the original appearance from the profile, we first pad the patches at each scale with the geometry configured by $\boldsymbol{s}$ to recover the individual channels. As the scales are complementary, all channels are then accumulated to recover the

[3] Spectrum cropping as image downsampling is explained in Appendix D and at `https://goo.gl/dht58Q`.

object appearance, see Fig. 4.1(g).

4.2 WAP Fitting with Supervised Descent Method

We introduce an implicit WAP fitting algorithm in this section. We deduce the true shape $\boldsymbol{s}^*$ from the observation at an initial shape $\mathcal{A}(\boldsymbol{s}_0)$, which is to solve a regression problem, $\mathcal{A}(\boldsymbol{s}^{(0)}) \mapsto \boldsymbol{s}^*$. The SDM algorithm [11] is adopted to solve the mapping and regression function. A major difference between SDM and a conventional discriminative method to solve a regression problem, is that the conventional methods only use one step regression, which may lead to lower performance. More recent work on boosted regression [101, 102, 103, 129] learns a set of weak regressors to model the mapping function. SDM is developed to solve general non-linear Least Square problems while boosted regression is a greedy method to approximate the mapping function. In the original gradient boosting formulation [100], feature vectors are fixed throughout the optimization, while [103, 129] re-sample the features at the updated landmarks for training different weak regressors. Although they have shown improvements using those re-sampled features, feature re-generation in regression is not well understood and invalidates some properties of gradient boosting. In SDM, the linear regressor and feature re-generation come up naturally from Newton's method. It is illustrated in [11] that a Newton update can be expressed as a linear combination of the feature differences between the one extracted at current landmark locations and the template.

Specifically, with SDM algorithm the mapping between the shape and local feature observation can be decomposed into a set of regressors and fitted recursively,

$$\begin{cases} \mathcal{A}(\boldsymbol{s}^{(i)}) \mapsto \Delta\boldsymbol{s}^{(i)}, \\ \boldsymbol{s}^{(i+1)} = \boldsymbol{s}^{(i)} + \Delta\boldsymbol{s}^{(i)}. \end{cases} \tag{4.2}$$

Each regressor is modelled linearly which gives,

$$\Delta \boldsymbol{s}^{(i)} = R^{(i)} \mathcal{A}(\boldsymbol{s}^{(i)}) + \boldsymbol{b}^{(i)}. \tag{4.3}$$

The parameters $\{R^{(i)}, \boldsymbol{b}^{(i)}\}$ can be learnt from the training images. Specifically, at each iteration, the the parameters can be learnt by minimising the residual error of regression in the training set,

$$\underset{\{R^{(i)}, \boldsymbol{b}^{(i)}\}}{\arg\min} \sum_{k=1}^{M} ||\Delta \boldsymbol{s}_k^{(i)} - R^{(i)} \mathcal{A}_k(\boldsymbol{s}_k^{(i)}) - \boldsymbol{b}^{(i)}||_2^2, \tag{4.4}$$

in which M is the number of training samples. $\Delta \boldsymbol{s}_k^{(i)}$ is the difference between the current shape $\boldsymbol{s}^{(i)}$ and the true shape $\boldsymbol{s}_k^*$ of the k-th training data. In all cases the initial shape $\boldsymbol{s}^{(0)}$ for the first regressor is set as the average shape at the average location in the training dataset. The shape samples for training the subsequent regressors are generated by applying the previous regressor,

$$\boldsymbol{s}_k^{(i+1)} = \boldsymbol{s}_k^{(i)} + R^{(i)} \mathcal{A}_k(\boldsymbol{s}_k^{(i)}) + \boldsymbol{b}^{(i)}, \tag{4.5}$$

In practice, to suppress the over-fitting problem with high-dimensional features and inadequate training data, an L2 regularisation is applied and the objective function (4.4) becomes,

$$\underset{\{R^{(i)}, \boldsymbol{b}^{(i)}\}}{\arg\min} \sum_{k=1}^{M} ||\Delta \boldsymbol{s}_k^{(i)} - R^{(i)} \mathcal{A}_k(\boldsymbol{s}_k^{(i)}) - \boldsymbol{b}^{(i)}||_2^2 + \lambda ||R^{(i)}||_2^2, \tag{4.6}$$

where λ controls the extent of regularisation. More details of SDM can be found at [11, 108].

To reduce the dimensionality of the descriptors and enhance the fitting performance, instead of using intensity features, a more robust feature descriptor such as HOG can be readily applied on the patches. Denote $h(\cdot)$ as the feature extraction

function, the fitting process can be expressed by,

$$\begin{cases} \Delta \boldsymbol{s}^{(i)} = R^{(i)} h(\mathcal{A}(\boldsymbol{s}^{(i)})) + \boldsymbol{b}^{(i)}. \\ \boldsymbol{s}^{(i+1)} = \boldsymbol{s}^{(i)} + \Delta \boldsymbol{s}^{(i)}, \end{cases} \tag{4.7}$$

with the parameters $\{R^{(i)}, \boldsymbol{b}^{(i)}\}$ learnt in the training data by,

$$\underset{\{R^{(i)}, \boldsymbol{b}^{(i)}\}}{\arg\min} \sum_{k=1}^{M} ||\Delta \boldsymbol{s}_k^{(i)} - R^{(i)} h(\mathcal{A}_k(\boldsymbol{s}_k^{(i)})) - \boldsymbol{b}^{(i)}||_2^2 + \lambda ||R^{(i)}||_2^2, \tag{4.8}$$

4.3 Pathology Classification

For the classification tasks, the correspondence of anatomical features should be built such that each entry of the descriptors corresponds to the same anatomical feature across the cases. The differences among the descriptors would therefore account for the true variations rather than the miss-alignment. In a WAP the appearance correspondence is built by extracting local features at corresponding landmarks. A classifier predicts the label ℓ given an anatomical observation $\Phi = \{\mathcal{A}, \boldsymbol{s}\}$, i.e, $\ell = \arg\max p(\ell|\Phi)$. The most significant variations in the training data $\{\Phi\}_i$ can be learned by principal components analysis and the dimensionality reduced by preserving the first t significant components, which span a feature space $P \in \mathbb{R}^{M \times t}$ with M being the dimensionality of Φ. A WAP therefore can be represented in the feature space by a compact set of parameters $\boldsymbol{b}_\Phi$, i.e., $\boldsymbol{b}_\Phi = P^T(\Phi - \bar{\Phi})$, in which $\bar{\Phi}$ is the mean of $\{\Phi\}_i$. Using $\boldsymbol{b}_\Phi$ as inputs the classifier now predicts $\ell = \arg\max p(\ell|\boldsymbol{b}_\Phi)$. We train the classifier with the AdaBoost method, with decision trees as the weak learners. A flowchart of training and applying the appearance models for landmark detection and pathology classification is given in Fig. 4.3.

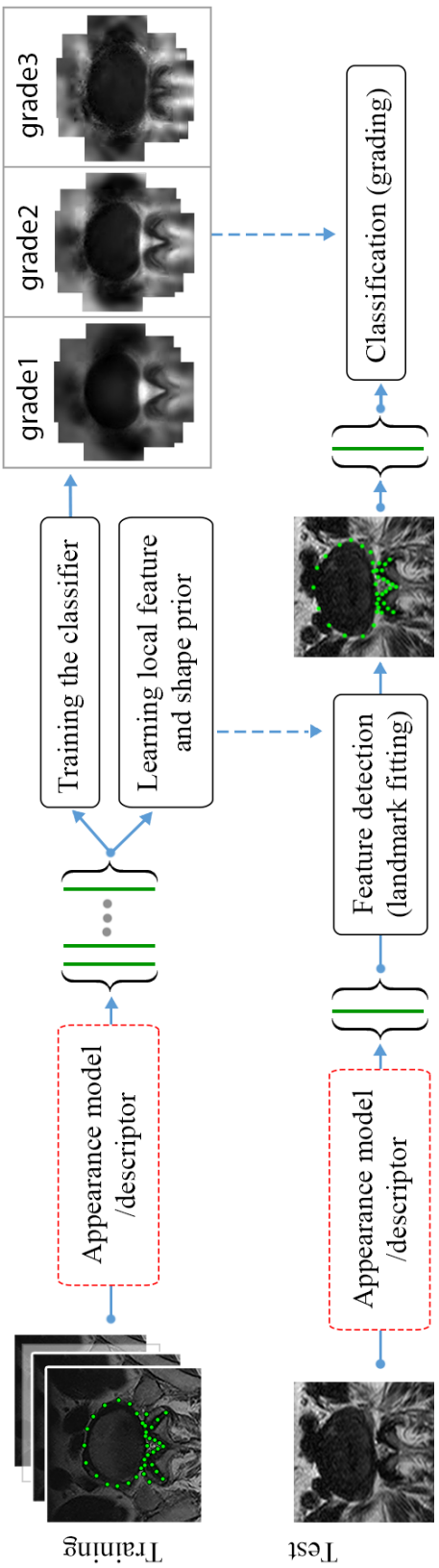


Figure 4.3: Pipeline of landmark detection and pathology classification.

Table 4.1: Performance of landmark detection

Metrics	AAM	ASM	CLM	DAP	WAP
PtoBD (in pixels)	3.10 ± 1.29	2.51 ± 1.32	2.34 ± 1.15	2.21 ± 1.07	**1.87±0.73**
DSC (%)	90.6 ± 4.9	92.1 ± 5.2	92.4 ± 5.2	92.8 ± 4.0	**94.7±2.6**

4.4 Results and Discussion

4.4.1 Landmark detection

To increase the scale of the dataset and cover richer pathological variations, we perform the landmark detection on the mixed dataset containing all 600 lumbar vertebral images. We run a single two-fold cross validation by randomly choose 300 images for training and detect the landmarks on the remaining 300. We compare the proposed WAP with AAMs, ASMs, CLMs, as well as Gaussian version DAPs proposed in Chapter 3 in order to validate the improvement of the loglet pyramid decomposition. The mean results of landmark detection are shown in Table 4.1. We can see that the WAP outperforms the other methods by a favourable margin. The lower variances also indicate its superior performance in terms of consistency.

4.4.2 Pathology classification

We apply the DAP for classifying two conditions, namely central canal stenosis and foraminal stenosis. For central stenosis, in each of the three subsets, the morphology of the central canal is inspected and labelled with three grades: normal, moderate and severe. For foraminal stenosis each case is annotated by the normal/abnormal labels. We randomly pick 100 samples to train the classifier and test on the remaining 100, and repeat for 100 times for an unbiased results.

The WAP extracted from the detected landmarks are projected onto the feature space and represented by a compact set of parameters $\boldsymbol{b}_\Phi$, which are used as inputs of the classifier. The average appearances delineated by WAPs are given in Fig. 4.4. The performance of normal/abnormal classification is measured with

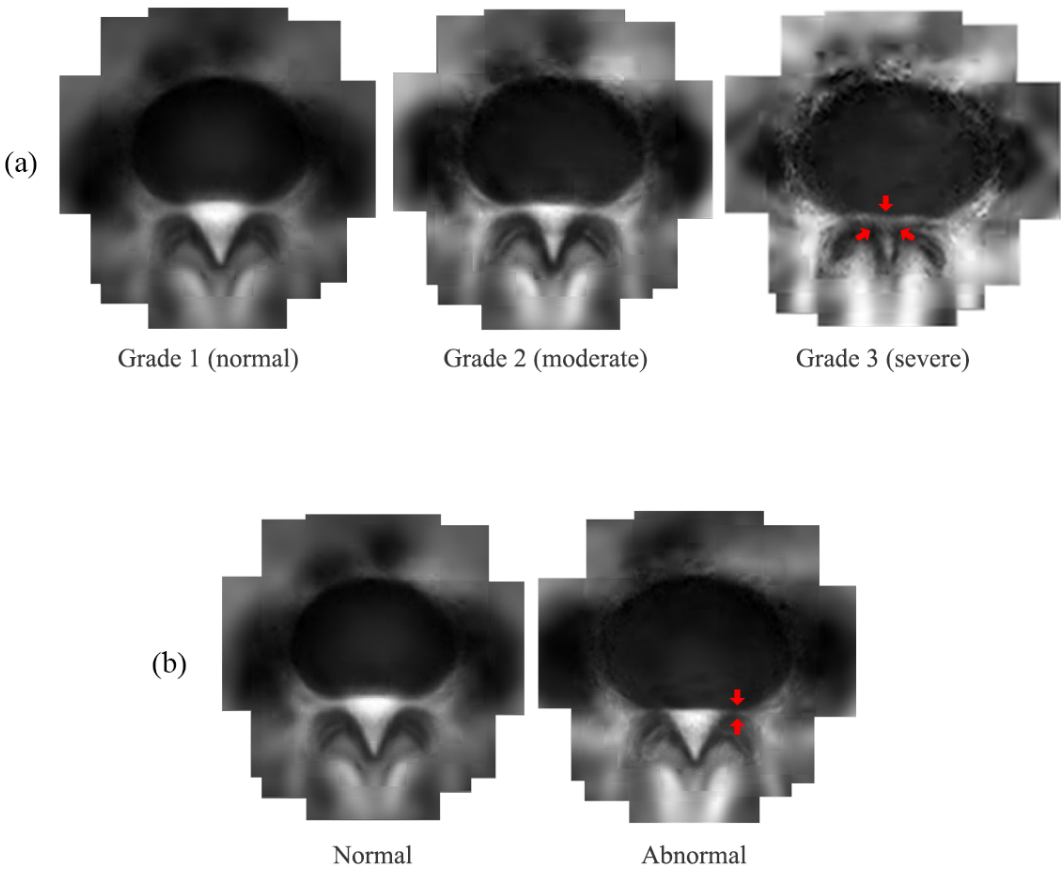


Figure 4.4: Average appearance of classes represented by WAP. (a) Three grades of central stenosis. (b) Normal and abnormal in terms of foreminal stenosis.

accuracy, which is calculated by $(TP+TN)/(TP+TN+FP+FN)$. The grading errors are measured with Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). We compare the performance of our method against approaches using other models as inputs to the same classifier. The agreements of the results with manual inspection are reported in Table. 4.2. Note that to keep the evaluation simple, we treat the misgrading between grade one and two the same as between grade two and three, the error of which are both one in grade. In the meanwhile the mislabelling between the normal and abnormal cases are evaluated with the accuracy of the two-class classification. Similarly we perform another normal/abnormal

Table 4.2: Agreement of classification and grading of central stenosis

Method	Accuracy (%) of classification			MAE of grading			RMSE of grading		
	L3/4	L4/5	L5/S1	L3/4	L4/5	L5/S1	L3/4	L4/5	L5/S1
ASM	79.1±4.8	77.4±4.3	81.7±4.5	0.25	0.31	0.20	0.55	0.67	0.48
AAM	70.1±7.1	69.7±7.3	71.3±8.8	0.41	0.44	0.32	0.72	0.79	0.58
CLM	81.0±4.9	82.4±4.5	82.7±4.4	0.23	0.25	0.23	0.53	0.56	0.52
DAP	80.7±4.9	82.1±4.6	84.7±4.2	0.23	0.25	0.18	0.53	0.58	0.47
WAP	**84.7±4.6**	**84.5±4.3**	**85.9±4.2**	**0.19**	**0.21**	**0.16**	**0.48**	**0.54**	**0.44**

Table 4.3: Accuracy (%) of classification of foreminal stenosis

Anatomy	ASM	AAM	CLM	DAP	WAP
L3/4	83.3 ± 3.8	73.3 ± 5.5	83.1 ± 4.7	84.3 ± 4.1	**85.0±3.9**
L4/5	82.4 ± 4.6	76.2 ± 5.8	83.3 ± 4.3	86.9 ± 3.9	**87.8±3.5**
L5/S1	81.8 ± 4.7	74.5 ± 5.7	82.9 ± 4.5	85.2 ± 4.3	**85.7±4.3**

classification on the morphology of the neural foremen. The classification accuracy of methods compared is reported in Table. 4.3. The results show that the Gaussian DAPs achieve better performance in central stenosis classification and grading comparing with ASMs, AAMs, and competitive performance with CLMs. For the task of foraminal stenosis classification the DAPs outperform the conventional models. A further improvement is observed in WAP models, which outperform other methods by a favourable margin.

A possible reason why the improvement by the Gaussian DAPs in central stenosis classification is not significant is that the larger size patches at lower levels of an DAP contain complex background textures, and therefore are not able to generalise well for large deformation. The improvement in WAPs therefore validates the argument that decomposing the complex textures into simpler wavelet components leads to better classification performance.

4.5 Summary

In this chapter we presented a novel WAP appearance model combining the DAP and wavelets. The WAPs were applied to detecting the landmarks and classifying

the LSS pathologies. The wavelet decomposition improves the performance in both tasks. In Appendix A, we further investigated the properties of wavelet decomposition, extended the wavelet local feature description to the face alignment scenario and validated its performance.

CHAPTER 5

Weakly Supervised Evidence Pinpointing

Towards Large Scale Learning on Weakly Annotated Data

In a standard clinical routine, a clinician often inspects consistent and salient structures for localising anatomies from radiological images, then evaluates the appearance of certain local regions for evidence of pathology. In a computer-aided approach, by learning to identify or *pinpoint* these regions and describing them discriminatively could provide precise information for localising and classifying the anatomies. In this chapter, we describe a method to automatically pinpoint the evidence regions as well as learn the discriminative descriptors in a weakly-supervised manner, i.e., only the class labels are used in training, and no other supervisory information is required. For localisation, we learn which features describe saliently the anatomies on a training set of aligned images. For classification, given the images with pathological labels, we learn the local features which provide evidence for discriminating between the normal and abnormal cases. We interpret evidence region pinpointing as a sparse descriptor learning [17, 18] problem in which the optimal feature descriptors are selected from a large candidate pool with various locations

and sizes. Because of its large scale, the problem is formulated in a stochastic learning manner and the regularised dual averaging (RDA) [19, 20] algorithm is used for the optimisation for several reasons: (1) The objection function of our method is composed as a convex function including two convex terms: one is the loss function of the learning task, and the other is a simple regularisation term. The RDA is designed and proved to be an efficient algorithm for convex optimisation; (2) We seek for sparse solutions such that the most relevant evidence regions can be selected. The RDA satisfies the requirement for promoting sparsity; (3) The stochastic learning and online optimisation manner of RDA is suitable for learning on large scale dataset with limited computing resource.

The learnt descriptors have several advantages over conventional hand-crafted representations, such as shape and appearance models, and local features, e.g., HOG [22] and local binary patterns [23]: (1) The training is weakly-supervised requiring no annotation of key features; (2) The learnt descriptors are more discriminative and informative, and therefore can contribute to better localisation and classification performance; (3) The evidence regions supporting the classification are automatically pinpointed which may be used by clinicians to determine the aetiology.

As already noted in Chapter 2, a CNN architecture [24, 25, 26, 27] can be trained to discriminative features on pathological labels with weak supervision as well, but requires large number of training samples and sufficient training. Instead of learning from raw image pixels, we formulate it as salient feature learning from a higher-level description of the image, which circumvents the low-level feature training. As a result the optimisation is straightforward consuming much less computing resource, and requiring no massive training data and no parameter tuning. Moreover, our descriptor learning method differs from the recent CNN based evidence pinpointing techniques [28, 29] in that we not only localise the evidence regions but at the same time give the description of these regions at optimal feature scales.

We show that compared with supervised methods trained with labels and landmarks, the descriptor learning method gives competitive performance trained on the same subsets with labels only. With further training on the weakly annotated subset, a significant improvement is obtained which validates the learning ability of the proposed method with weak-supervision.

5.1 Methodology

An anatomy can be localised by certain salient local structures distinctive from the background. Also, a pathological condition in an anatomy is often shown as changes in intensity or structure in local regions. Learning to identify and describe these discriminative regions accordingly can therefore capture the key information for localisation and classification tasks. We detail the formulation and optimisation of discriminative region learning below.

5.1.1 Formulation

Assume we have a set of training images classified into a subset $\mathcal{N}$ with negative labels and a subset $\mathcal{P}$ with positive labels. For example, for classification tasks $\mathcal{N}$ and $\mathcal{P}$ consist of normal and pathological images. For localisation tasks, $\mathcal{N}$ refers to the images with the anatomies aligned, and $\mathcal{P}$ the misaligned images.

To learn the local regions and features that lead to the classification, we generate a pool of region candidates having various locations and scales, and select the most discriminative ones. Specifically, to generate the location candidates, each region is represented by a Gaussian weighted window $g(\rho, \theta, \sigma)$ with ρ and θ being the polar coordinate of the window on the image, and σ the size of the window, see Fig. 5.1(a). Parameters $\{\rho, \theta\}$ are sampled over the ranges $\rho = [0, \rho_1]$, $\theta = [0, 2\pi]$ such that the regions cover the whole image. To include multiple sizes of local features in the candidate pool, we build an image pyramid with the lower resolution

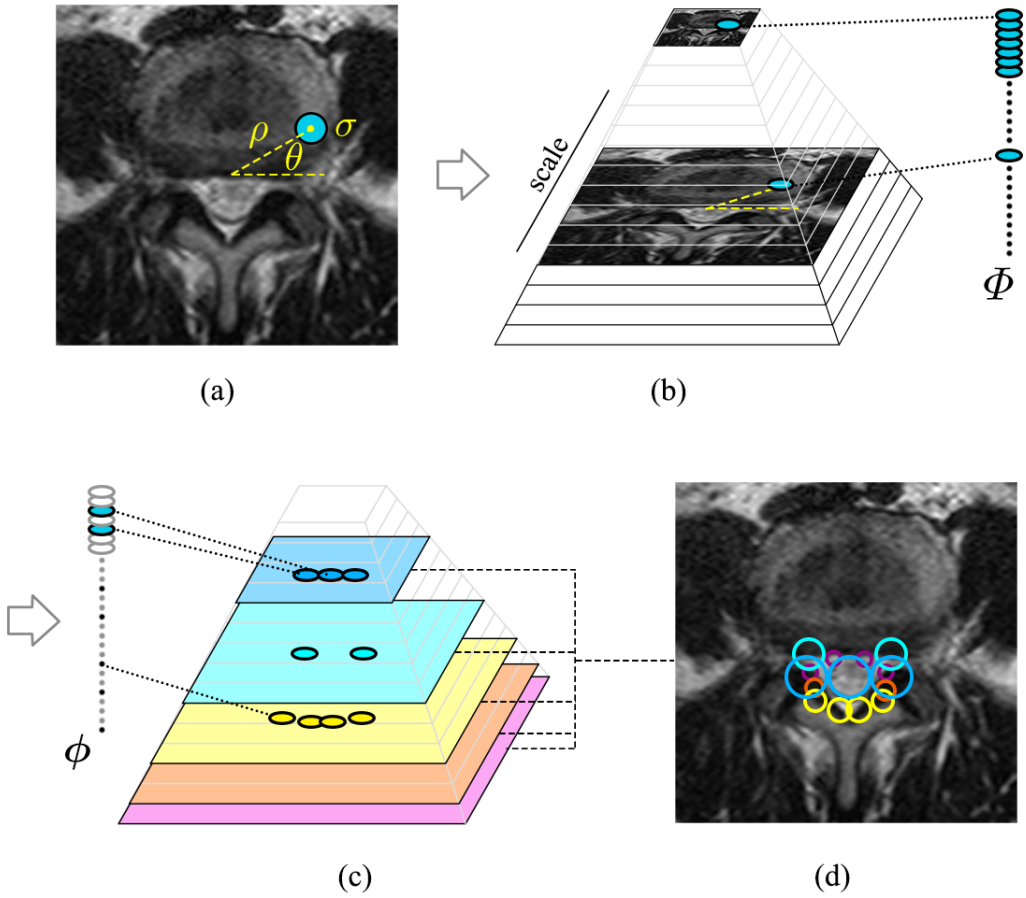


Figure 5.1: (a) Region candidate on an image. (b) Region candidates on an image pyramid to include multiple region sizes and feature scales. (c) For a certain task, the salient regions are selected by sparse learning. (d) The learnt descriptors.

images containing larger scale textures. The region candidates are sampled from each layer of the pyramid with the same size in pixels, which results in larger effective region sizes and feature scales on lower resolution images, see Fig. 5.1(b).

To represent each region, instead of using raw image features, we decompose the local textures into complementary frequency components for a compact description. This is achieved by designing window functions to partition the spectrum, see Fig 5.2(a). The specific form of the windows are shown in Fig 5.2(b). The low-pass window is a Gaussian function, and the oriented windows are logarithmic functions along radius in four directions. Each of the 4 oriented windows in Fig 5.2(b) corre-

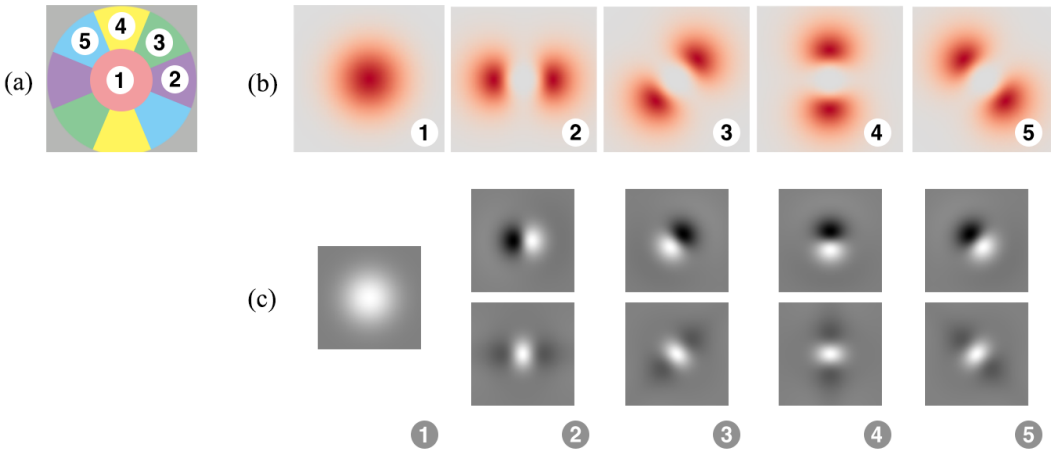


Figure 5.2: (a) Spectrum partition. (b) Filter windows in the Fourier domain. (c) Filters in the spatial domain corresponding to intensity, gradient and curvature features respectively.

sponds to 2 spatial-domain filters (real and imaginary part separately) and together with the low-pass filter, we obtain 9 filters, see Fig 5.2(c). Note that the filters correspond to the intensity, first and second order derivative features respectively. The filters are similar to Haar and discrete wavelet filters but with enhanced smoothness and complementary properties. We calculate the response map of the image to each filter, and accumulate over the i-th region to obtain the region descriptor $\Phi_i \in \mathbb{R}^{1\times 9}$. The region descriptors from different locations and pyramid levels form a candidate pool $\Phi = \{\Phi_i\}_{i=1}^{N} \in \mathbb{R}^{N\times 9}$, where N is the total number of the region candidates. Φ gives a redundant (overcomplete) description of the image, see Fig. 5.1(b).

The task then is to select from the candidate pool Φ a few regions containing the discriminative information, which we formulate as a sparse learning problem. The selection can be described by the operation,

$$\boldsymbol{\phi} = W^{\frac{1}{2}}\Phi. \tag{5.1}$$

$W \in \mathbb{R}^{N\times N}$ is a diagonal matrix with sparse entries $\boldsymbol{w} = [w_1, w_2, \ldots, w_N]$, in which w_i is the assigned weight of the i-th region Φ_i, and the non-zeros weights correspond-

ing to the regions selected. $\boldsymbol{\phi}$ represents the selected salient features (Fig. 5.1(c)).

The objective is to learn $\boldsymbol{w}$ such that the selected descriptors $\boldsymbol{\phi}$ are consistent within class and discriminative between classes. Let $\boldsymbol{\phi}(p)$, $p \in \mathcal{P}$ and $\boldsymbol{\phi}(n)$, $n \in \mathcal{N}$ be the descriptors of two random examples from the positive and negative image set respectively. The distances between the descriptors can be calculated by,

$$||\boldsymbol{\phi}(p) - \boldsymbol{\phi}(n)||_2^2 = \sum_{i=1}^{N} ||\sqrt{w}_i \Phi_i(p) - \sqrt{w}_i \Phi_i(n)||^2 = \sum_{i=1}^{N} w_i ||\Phi_i(p) - \Phi_i(n)||^2 = \boldsymbol{w}^T \boldsymbol{d}(p,n), \tag{5.2}$$

where $\boldsymbol{d}(p,n) \in \mathbb{R}^{N \times 1}$ is a vector with each entry $d_i(p,n)$ being the feature difference calculated at a region, i.e., $d_i(p,n) = ||\Phi_i(p) - \Phi_i(n)||^2$.

Similarly we randomly sample two examples $n_1, n_2 \in \mathcal{N}$ from the negative set and calculate the distance denoted by $\boldsymbol{d}(n_1, n_2)$. To penalise the differences within the negative set and reward the distances between the positive and negative sets, we set a margin-based constraint,

$$\boldsymbol{w}^T \boldsymbol{d}(n_1, n_2) + 1 < \boldsymbol{w}^T \boldsymbol{d}(p,n). \tag{5.3}$$

We do not penalise the differences within the positive set as it represents the misaligned or pathological images with large variations, see Fig. 5.4(a).

The objective function enforcing the constraint may be composed in a sparse learning form,

$$\underset{\boldsymbol{w} \geqslant 0}{\arg\min} \sum_{p \in \mathcal{P}; n, n_1, n_2 \in \mathcal{N}} \mathcal{L}(\boldsymbol{w}^T \boldsymbol{d}(n_1, n_2) - \boldsymbol{w}^T \boldsymbol{d}(p,n)) + \mu ||\boldsymbol{w}||_1, \tag{5.4}$$

where $\mathcal{L}(z) = \max\{z + 1, 0\}$ is a loss function penalising the non-discriminative entries, and the $\ell 1$-norm $||\boldsymbol{w}||_1$ is a sparsity-inducing regulariser which encourages the entries of $\boldsymbol{w}$ to be zero, thus performs region selection. The 'sparsity' here means that we want to select a few most relevant regions from a large candidate pool. Note that each n in the function represent an independent random index from the negative

set, and p from the positive set. The number of the summands is not fixed, which fits with the stochastic learning and online optimisation procedure, i.e., repetitively drawing random samples $\boldsymbol{d}(n_1, n_2)$, $\boldsymbol{d}(p, n)$ and optimising $\boldsymbol{w}$ until a criterion is met. The random sampling also enables incremental learning which means we can refine the model without re-learning it all over again when new training data become available. We deduce the solution to (5.4) in the next section.

5.1.2 Optimisation

Finding the sparse parameter $\boldsymbol{w}$ in (5.4) is a regularised stochastic learning problem where the objective function is the sum of two convex terms: one is the loss function of the learning task fed recursively by random examples, and the other is a $\ell 1$-norm regularisation term for promoting sparsity. It can be solved efficiently by the RDA algorithm [19, 20], which recursively learns and updates $\boldsymbol{w}$ with new examples.

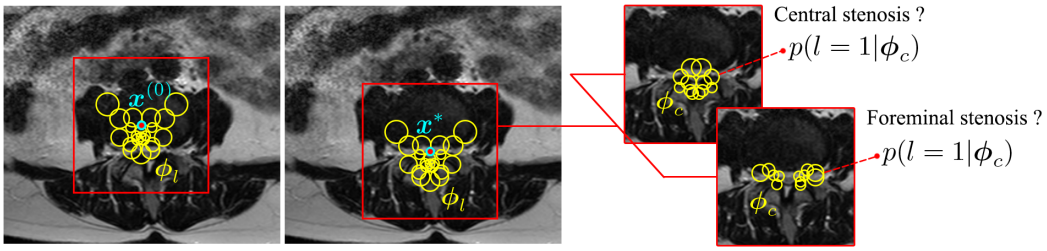
At the t-th iteration, RDA takes in a new observation, which in our case are random pairs $\boldsymbol{d}(p, n)$ and $\boldsymbol{d}(n_1, n_2)$. The loss subgradient $\boldsymbol{g}_t$ is calculated by,

$$\begin{aligned}\boldsymbol{g}_t =& \frac{\partial \mathcal{L}\left(\boldsymbol{w}^T\left(\boldsymbol{d}(n_1, n_2) - \boldsymbol{d}(p, n)\right)\right)}{\partial \boldsymbol{w}} \\ =& \begin{cases} \boldsymbol{d}(n_1, n_2) - \boldsymbol{d}(p, n), & \boldsymbol{w}^T\left(\boldsymbol{d}(n_1, n_2) - \boldsymbol{d}(p, n)\right) > -1 \\ 0, & otherwise. \end{cases}\end{aligned} \tag{5.5}$$

$\boldsymbol{g}_t$ is used to update the average subgradient, $\bar{\boldsymbol{g}}_t = \frac{1}{t}\sum_{i=1}^{t} \boldsymbol{g}_i$. Updating the parameter $\boldsymbol{w}$ with RDA takes the form,

$$\boldsymbol{w}_{t+1} = \arg\min_{\boldsymbol{w}}(\boldsymbol{w}^T \bar{\boldsymbol{g}}_t + u||\boldsymbol{w}||_1 + \frac{\beta_t}{t} h(\boldsymbol{w})) \tag{5.6}$$

in which the last term is an additional strong convex regularisation term. One can set $h(\boldsymbol{w}) = \frac{1}{2}||\boldsymbol{w}||_2^2 = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$, $\beta_t = \gamma\sqrt{t}, \gamma > 0$ for a convergence rate of $O(1/\sqrt{t})$. By writing $\boldsymbol{u}$ as a N dimension vector with each elements being u, equation (5.6)



(a) Anatomy localisation (b) Task-specific classification

Figure 5.3: Applying the learnt descriptors for localisation and classification.

becomes,

$$\boldsymbol{w}_{t+1} = \arg\min_{\boldsymbol{w}}(\boldsymbol{w}^T\bar{\boldsymbol{g}}_t + \boldsymbol{w}^T\boldsymbol{u} + \frac{\gamma}{2\sqrt{t}}\boldsymbol{w}^T\boldsymbol{w}), \tag{5.7}$$

which can be solved by Least Squares method to give,

$$\boldsymbol{w}_{t+1} = -\frac{\sqrt{t}}{r}(\bar{\boldsymbol{g}}_t + \boldsymbol{u}). \tag{5.8}$$

The discriminative regions and optimal descriptors are obtained by keeping only the candidates with non-zero weights indicated by the learnt $\boldsymbol{w}$. An example is given in Fig. 5.1(d).

5.1.3 Localisation and classification

Denoting $\boldsymbol{\phi}_l$ as the learnt optimal descriptor for localising anatomy, and $\boldsymbol{\phi}_c$ the descriptor for a certain classification task, we show how the optimal descriptors are applied (Fig. 5.3).

Localisation. The anatomy is described discriminatively by $\boldsymbol{\phi}_l$ which represents the salient structures. Localising the anatomy in the image is conducted by searching for these structures. Given an initial estimation $\boldsymbol{x}^{(0)}$ of the location, which can be set at the centre of the image, the descriptor at the initial location $\boldsymbol{\phi}_l(\boldsymbol{x}^{(0)})$ is observed to deduce the true location $\boldsymbol{x}^*$. The deduction can be expressed as solving the regression $\boldsymbol{\phi}_l(\boldsymbol{x}^{(0)}) \mapsto \boldsymbol{x}^*$. The direct mapping function is non-linear in nature

and training such function comes up against the over-fitting problem. In practice the mapping can be decomposed into a sequence of linear mapping and updating steps,

$$\begin{cases} \text{Mapping:} & \boldsymbol{\phi}_l(\boldsymbol{x}^{(k)}) \mapsto \Delta\boldsymbol{x}^{(k)}, \\ \text{Updating:} & \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \Delta\boldsymbol{x}^{(k)}, \end{cases} \tag{5.9}$$

where in the mapping stage, a prediction for the correction of the location is made, based on the observation $\boldsymbol{\phi}_l(\boldsymbol{x}^{(k)})$ at the current location $\boldsymbol{x}^{(k)}$; and in the updating stage, the location and observation is updated. The learning mapping function is set to be,

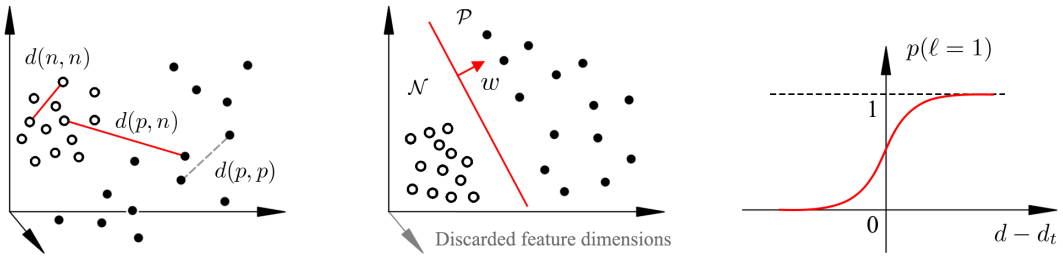
$$\Delta\boldsymbol{x}^{(k)} = R^{(k)}\boldsymbol{\phi}_l(\boldsymbol{x}^{(k)}) + \boldsymbol{b}^{(k)}, \tag{5.10}$$

with $R^{(k)}$ being a projection matrix and $\boldsymbol{b}^{(k)}$ the bias. $\{R^{(k)}, \boldsymbol{b}^{(k)}\}$ in each iteration is trained with the Supervised Descent Method, the details of which can be found in [11], and Section 4.2.

Classification. Learning $\boldsymbol{w}$ in the objective function (5.4) can be viewed as a simultaneous feature selection and classification process. The zero entries in $\boldsymbol{w}$ correspond to the non-salient features (or region candidates) to be discarded. In fact, the non-zero entries in $\boldsymbol{w}$ form a vector defining the hyperplane classifying the positive and negative samples in the salient feature space, which is similar to a support vector in Support Vector Machine classifier, see Fig 5.4(b).

For a specific pathological condition, the learnt descriptor $\boldsymbol{\phi}_c$ covers the regions where the abnormalities are most likely to appear, and preserves their discriminative features for classification. To predict the class label ℓ of a test image, we extract the descriptor $\boldsymbol{\phi}_c(\boldsymbol{x}^*)$ at the detected location $\boldsymbol{x}^*$ and calculate the average distance to the normal descriptors,

$$d = \frac{1}{|\mathcal{N}|}\sum_{n\in\mathcal{N}} ||\boldsymbol{\phi}_c(\boldsymbol{x}^*) - \boldsymbol{\phi}_c(n)||_2^2, \tag{5.11}$$



(a) Objective (b) Classification (c) Probability

Figure 5.4: (a) The objective of sparse descriptor learning. (b) The zero entries in the learnt $\boldsymbol{w}$ remove the non-salient feature dimensions (region candidates), the non-zero entries define the hyperplane for classification in the salient feature space. (c) The sigmoid probability function.

where n indexes all the cases in the normal set $\mathcal{N}$.

A larger d indicates a greater probability of the case being abnormal. More formally, the probability of the case being abnormal is modelled by a sigmoid function (Fig. 5.4(c)),

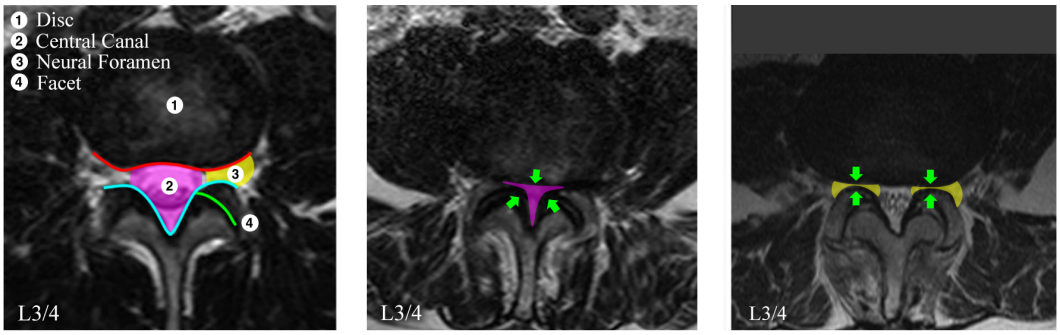
$$p(\ell = 1|\boldsymbol{\phi}_c(\boldsymbol{x}^*)) = \frac{1}{1 + e^{-(d-d_t)}}, \tag{5.12}$$

where d_t is a threshold distance. The cases with $p > 0.5$ are classified as abnormal, with confidence p. Conversely, the cases with $p < 0.5$ are classified as normal with the confidence $(1 - p)$.

5.2 Experiments

5.2.1 Validation protocols.

We train and validate the weakly supervised evidence pinpointing technique on both the three densely annotated subsets LSS containing around 200 cases and the three weakly annotated datasets containing 600 cases. In each of the three intervertebral subsets we randomly select 100 densely annotated images as the test set, and the remaining densely annotated images as the training set. The landmarks and pathology labels are used for training the conventional supervised methods. In contrast the



(a) Anatomy (b) Severe central stenosis (c) Foraminal stenosis

Figure 5.5: (a) Anatomy of a disc-level intervertebral slice. (b) A case with severe central stenosis. (c) A case with foraminal stenosis.

evidence pinpointing method learns on the pathology labels requiring no landmarks. The additional images with only classification labels are used for further training the evidence pinpointing method to validate its learning ability on weakly annotated large scale data. The selection of training and test set is repeated for an unbiased validation. The training set is used for learning descriptors for both localisation and task-specific classifications. In the testing stage, the localisation and classification tasks are conducted by each method independently, and the performance is evaluated.

5.2.2 Anatomy localisation.

The learning result of the optimal descriptor for localising the vertebrae, L3/4 as an example, is shown in Fig. 5.6(a). The hot maps of salient regions are visualised by showing the selected region candidates as Gaussian blobs. It is interesting to compare these with the biological anatomy and medical definitions in Fig. 5.5 and the annotations by the clinician in Fig. 5.7(a). The learnt descriptor highlights the posterior margin of the disc and the posterior arch, which have sharp textures and high contrast. Note that compared with a clinician's annotations, the front edge of the disc is not selected. The reason for this may be there being less consistency across images because of the variation in disc size, as well as the ambiguous boundaries to

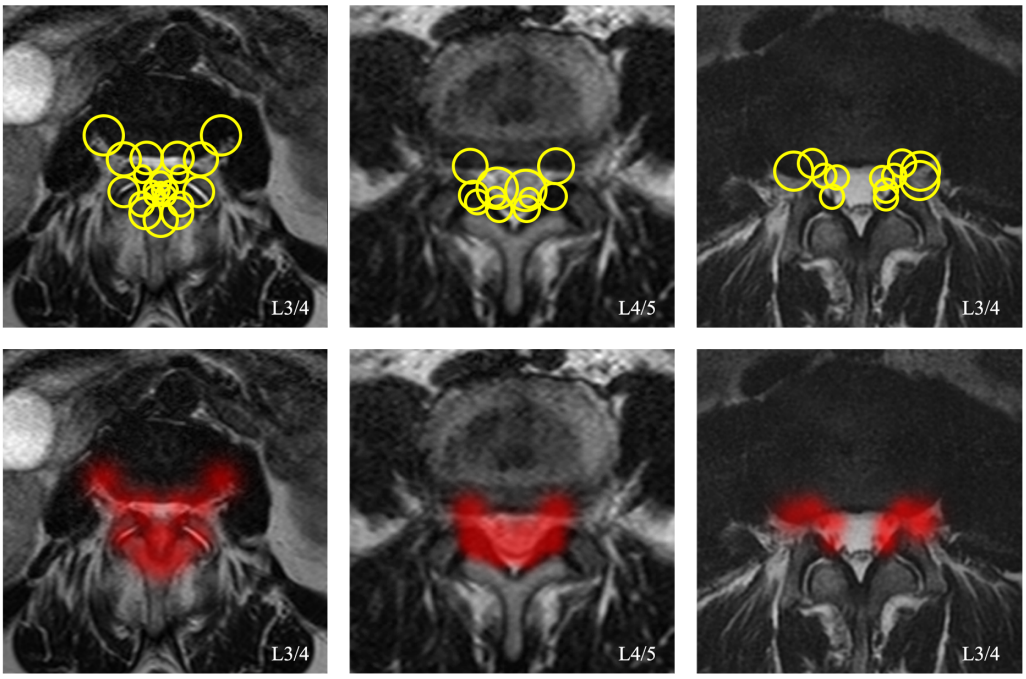
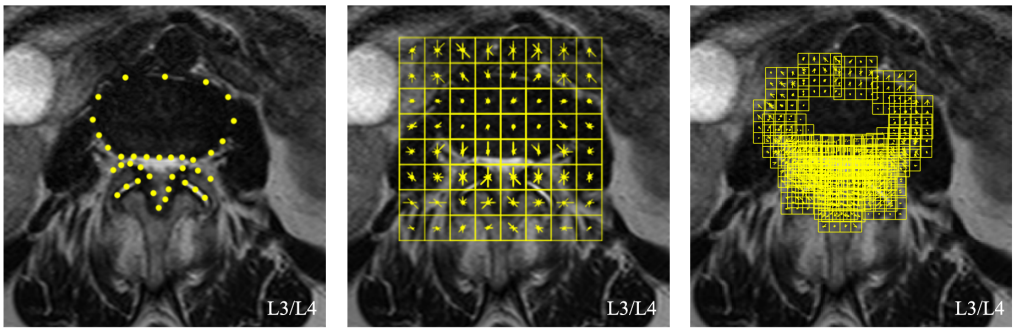


Figure 5.6: The discriminative descriptors (top) and evidence regions (bottom) learnt for the task of (a) anatomy localisation (b) central stenosis classification and (c) foraminal stenosis classification. The learnt evidence regions for classification show high consistency with the clinical definition of the conditions as shown in Fig. 5.5.

the abdominal structures in some of the cases.

We compare our method with HOG grid [22] and Deformable Part Models (DPM) [48, 23]. The HOG grid is a hand-crafted descriptor covering the holistic appearance, see Fig. 5.7(b). It assumes no prior clinical knowledge and assigns equal weights to the local features of the anatomy. The DPM is a strongly supervised method which describes the anatomy by local patches at each of the landmarks as well as the geometry of the landmark locations (Fig. 5.7(c)). Each patch is described by a SIFT descriptor. In all the methods the initial location is set at the centre of the images and the searching is driven by the SDM algorithm [11].

The experimental results are reported in Table 5.1. The average distances between the initial and the true locations are also given. We can see that our learnt



(a) Accurate annotation (b) HOG grid (c) DPM

Figure 5.7: Comparative descriptors: landmarks, HOG-grid and DPM.

Table 5.1: Precision of anatomy localisation in RMS error in mm ($^+$ Require landmarks. * Trained on additional weakly annotated data).

Data	Initial	HOG grid*	DPM$^+$	Learnt	Learnt*
L3/4	16.41±10.10	2.45 ± 1.69	2.01 ± 1.62	1.95 ± 1.58	1.22±1.01
L4/5	16.59±10.80	2.37 ± 1.55	1.73 ± 1.30	1.76 ± 1.26	1.57±1.36
L5/S1	12.86 ± 8.29	2.52 ± 1.71	1.85 ± 1.42	2.09 ± 1.52	1.24±0.96

descriptors give comparable localisation precision with DPM when trained on the same densely annotated subsets, but use no landmark annotation. With further training on additional data, a significant improvement is observed indicating the learning ability of our method on weakly annotated data. Note that the precision of localisation here is not comparable with the precision of landmark detection in Fig. 3.13 and Table. 4.1 because in landmark detection experiments the shape is initialised closely to the true location, and in anatomy localisation experiments here the location is initialised at the centre of the images, which is more challenging. Note also that the errors reported in Table 5.1 are in millimeters.

5.2.3 Pathology classification.

The classification follows on from the anatomy localisation step. The learnt discriminative descriptors and evidence regions for the classification of central canal stenosis and foramen stenosis are shown in Fig. 5.6(b)(c) respectively. We can see

that the descriptor learnt on central stenosis labels highlights the spinal canal area. When learnt on foraminal stenosis labels, it pinpoints the neural foramen as the evidence regions. These regions show high agreement with the relevant biological areas in Fig. 5.5(b).

The learnt descriptors are extracted at the detected location for classification on test images. The predicted pathological labels as well as the confidences of prediction are given by (5.12). For comparison, in the HOG grid method, the descriptors are centred at the detected location. The classifiers for the compared methods are trained with the AdaBoost method, with decision trees as the weak learners. The performance is evaluated by the agreement with labelling done by a clinician, calculated by $(pp + nn)/M$, in which pp and nn are the number of agreed positive and agreed negative cases, M is the total number of cases.

The results of the two classification tasks are shown in Table 5.2 and Table 5.3. We compare the descriptor learning method with the HOG grid method as well as DAPs and WAPs proposed in previous chapters. Note however the experiments set up for the descriptor learning is more challenging because of the three factors: Firstly the datasets used here are annotated with only class labels and no landmark; secondly the initial location is set at the centre of the images for all cases while for the DAPs and WAPs the initial locations are set pixels away from the true location; thirdly the features used in DAPs and WAPs are extracted after the landmark detection therefore the shape of the anatomy is already known.

Nevertheless the descriptor learning method gives better or competitive accuracies of central stenosis classification compared with supervised methods, trained on the same densely annotated subset but requires no landmarks to be identified. For the classification of foraminal stenosis, the descriptor learning method trained on large-scale weakly-annotated data shows comparable performance with the DAPs and WAPs. In both tasks, a significant improvement is seen with additional training on weakly annotated data validating the learning ability of the proposed methods

Table 5.2: Agreement (%) of classification of central canal stenosis. ($^+$ Require landmarks. * Trained on additional weakly annotated data.)

	Human	HOG grid	DAP$^+$	WAP$^+$	Learnt	Learnt*
L3/4	88.5	80.6 ± 4.9	80.7 ± 4.9	84.7 ± 4.6	85.7 ± 3.5	**87.2±3.2**
L4/5	87.4	81.3 ± 4.6	82.1 ± 4.6	84.5 ± 4.3	84.2 ± 3.4	**85.1±3.4**
L5/S1	89.2	81.8 ± 4.7	84.7 ± 4.2	85.9 ± 4.2	86.0 ± 3.7	**87.5±3.3**

Table 5.3: Agreement (%) of classification of foraminal stenosis. ($^+$ Require landmarks. * Trained on additional weakly annotated data.)

	Human	HOG grid	DAP$^+$	WAP$^+$	Learnt	Learnt*
L3/4	86.5	79.6 ± 4.5	84.3 ± 4.1	**85.0**±3.9	82.9 ± 4.5	84.3±3.9
L4/5	87.2	81.5 ± 4.9	86.9 ± 3.9	**87.8**±3.5	82.5 ± 4.5	84.0±4.0
L5/S1	89.5	81.7 ± 4.4	85.2 ± 4.3	85.7 ± 4.3	84.1 ± 3.8	**87.1±3.4**

with weakly supervision. Note also that the performance is affected by the precision of the human labels, as the clinician can only achieve a certain level of agreement with themselves when the labelling step is repeated on same dataset. We report the self-agreement of a clinician in Table 5.2 and Table 5.3, denoted as the human performance. The disagreement is generally caused by ambiguous conditions in many cases. We give several example images with different degrees of degenerations, and show the classification labels by the clinician as well as the labels and probabilities by our method in Fig. 5.8. The probability indicates the confidence of our prediction, which may be helpful for being aware of and understanding errors in the classification results.

5.3 Summary

In this chapter we propose a method for learning the optimal descriptors for the tasks of anatomy localisation and classification. The learnt descriptors for localising an anatomy highlights consistent and salient structures across a set of images. The descriptors for classifying a specific condition, learnt with no prior knowledge but only the labels, pinpoint the evidence regions where the abnormalities are most

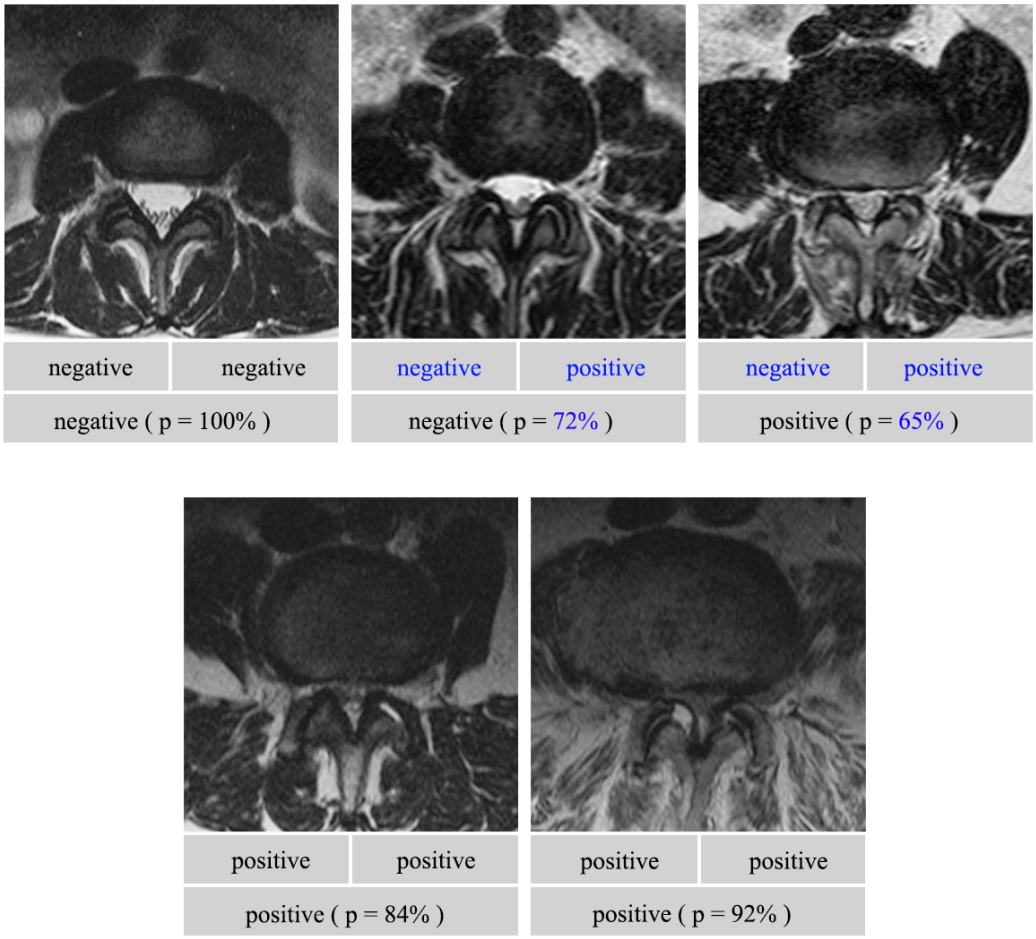


Figure 5.8: Example images with different degrees of degeneration. Results in the first row are the repeated labels for central stenosis by the same clinician, made at different times. Disagreement is shown in blue. Results in the second row are the labels and probabilities by our classification method.

likely to appear. The information in the descriptors is highly discriminative leading to more accurate classification results. The training is straightforward with no need of parameter tuning. We have shown that promising results are achieved when learnt on 600 weakly annotated images. The average training time for one task is about 27 minutes in MATLAB on a 3.20 GHz GPU with 16 GB RAM. The method can be readily applied to other clinical tasks for rapidly pinpointing and describing evidence of abnormalities directly from expertly labelled data.

CHAPTER 6

Conclusions

In this thesis we presented several novel methods for medical image analysis such as modelling anatomical appearance, localising anatomy and fitting landmarks, classifying pathologies, and pinpointing evidence.

We proposed the DAPs, a new class of appearance models, to delineate and represent the anatomies, and validated their performance in variation modelling and landmark fitting. We detailed how DAPs are constructed, trained and fitted and demonstrated that the appearance of an object can be delineated with multi-scale parts and that an associated deformation can be approximated by a set of locally rigid transformations of the parts. The DAP utilises the power of local feature searching and shape regularisation algorithms the same as a deformable part-based model. Meanwhile it enhances the robustness of part searching with multi-scale local feature descriptors. The prior knowledge of the anatomy variations is modelled and utilised with density estimation theory, through which the shape instances are regularised within plausible range. A subspace LK algorithm was derived for robust landmark detection.

We further proposed an extension to the DAP models referred to as WAPs. Here the images are decomposed into complementary wavelet channels, and the

appearance is decomposed into simple feature components. The SDM algorithm is employed to model the prior knowledge of anatomy variations implicitly, and to solve feature-to-shape regression for landmark detection and shape fitting. The WAP models were applied for landmark detection and pathology classification, and the improvement in performance was validated.

To be able to learn on large-scale weakly-labelled datasets, we proposed a new weakly supervised learning approach. The method learns how to describe the anatomy discriminatively. Salient and consistent features are used to localise the anatomy in the images, but also it learns to identify which specific region and features contribute to a certain classification, and reveal the pathologies supporting the diagnosis, a technique we call evidence pinpointing. The learning is weakly supervised which means only the pathological labels are required. We formulate evidence pinpointing as a sparse descriptor learning problem. To deal with the large size of the dataset, we compose the object function in a stochastic way and optimise it by the Regularised Due Averaging algorithm.

We validated the proposed methods on the lumbar vertebrae datasets for the problem of lumbar spinal stenosis. Comparing against AAMs in appearance modelling, the proposed DAPs preserve more precise textural information of the appearance. In landmark detection, the DAPs result in more accurate and robust performance when compared with conventional methods. As an extension to the DAPs, the WAPs were tested in the tasks of landmark detection and pathology classification. A improvement in performance is observed mainly benefiting from the complementary decomposition of image features by the wavelets. The evidence pinpointing method trained on large scale datasets is able to identify the region of interest related to a certain disease. The pinpointed regions show high consistency with the medical definition of the disease. The features extracted from the regions were then used for classifying the pathologies. Experiments results showed that the proposed weakly supervised method gives better or competitive performance in

anatomy localisation and pathology classification comparing against strongly supervised methods trained on densely labelled dataset.

In addition to applying the methods on lumbar spinal stenosis, we demonstrated the accuracy and efficiency of DAPs in 3D on the hip data for the problem of hip impingement. The 3D DAPs preserve more precise textural information of the appearance comparing the traditional AAMs, while consume much less memory. The computation is more efficient as well consuming less than 10 % the training time.

We also devote a chapter in Appendix 1 investigating the wavelet local feature pyramids in the WAP models, and integrating them with deformable part models. The performance is validated in the face detection context for the facial landmark detection. The proposed feature descriptor improves the performance of DPMs by a large margin, reporting state-of-the-art results on the popular HELEN dataset and competitive performance on 300-W dataset.

6.1 Future Research Directions

Future work may include investigating in the following aspects:

1. The proposed DAP as a general multi-scale part-based model could be integrated with other prior modelling and feature-to-shape regression algorithms in the machine learning literature. For example the pictorial structures [13] could be used to constrain the geometry of the parts in the models. The tree structure [52] could be employed to deal with large shape variations. The regularised mean-shift algorithm [14] could be adopted, with the salient map generated by the proposed LFPs.

2. To deal with further larger scale weakly labelled dataset or extend the methods to 3D, a random candidate sampling strategy can be introduced to the evidence pinpointing algorithm. Because each sampling step is independent in

random sampling and stochastic learning strategy, parallel computing could be used with the method in order to reduce the training time on large scale data. To deal with the increased scale in 3D problem, the candidate pool in the evidence region selection can itself be generated generate by sampling at random locations and scales in order to reduce the memory usage.

3. The proposed methods address the general tasks in medical image processing such as detection, segmentation, classification, evidence pinpointing, therefore they can be applied to other clinical tasks where similar medical image processing and understanding function is required. We plan to implement our methods on the knee to help the understanding of various conditions such as patellogemoral joint, medial and lateral tibial disease, and anterior cruciate ligament rupture. The DAPs and WAPs can be used to locate and segment the knee structures for geometrical reconstruction, and to classify the pathologies with strong supervision. The evidence pinpointing technique might potentially help identifying the pathological areas with respect to certain conditions.

4. CNN has recently emerged to be a popular technique to deal with large scale data. However the training of CNN is time-consuming and requires extensive computing resources. The first few layers in a CNN attempt to extract useful features from the images, introduce non-linearity in the network and reduce feature dimension. In our evidence pinpointing formulation, instead of learning from raw image pixels, we start from higher level description of images, which as a result speeds up the optimisation. In future work we may investigate into fitting our method into CNN architecture in order to learn on large scale clinical dataset more efficiently.

APPENDIX A

Wavelet Local Feature Pyramids for Part Description: An Extended Application to Face Alignment

In this chapter, to further study the properties and evaluate the performance of the local feature pyramid, we apply it in the computer vision field, e.g. face alignment, where the datasets and benchmarks are public available. We apply the local feature pyramids for facial part description in the Deformable Part Model (DPM) framework. Part descriptors in the DPM seek a representation of local structures which preserves intrinsic properties and discriminative information, while exhibiting invariance to changes such as illumination, scale, and variations in appearance across instances. The most successful part descriptors in DPMs are those based on oriented gradients such as SIFT [130]. The power of SIFT lies in its robustness to illumination and noise through neighbourhood pooling, and its invariance to scale achieved by salient scale selection. When SIFT descriptors are used as part 'experts' in DPMs, e.g., in [11, 131, 52], the scale is selected by assigning a patch size without salience detection, therefore salient local features may not be captured. We propose to integrate the loglets with SIFT to capture wider scale range therefore preserve

richer information in SIFT descriptors. SIFT descriptors accumulate the orientations of gradients in a local region, and calculate orientation histograms. We show how loglets can be converted to differential filters to generate oriented gradients with explicit scale selection, based on the fact that all differential filters take the form of imaginary odd-windows in the Fourier space. The new feature descriptor combines the pooling power of SIFT and scale selection of loglets and is therefore termed Loglet-SIFT (L-SIFT) [132].

Several original contributions are included in the proposed descriptor, namely,

1. We design differential filters directly in the Fourier domain with explicit scale selection;

2. A high pass oriented filter is generated by accumulating a group of adjacent loglets, which achieves a uniform coverage towards the Nyquist frequency and is able to preserve the sharpest gradients without aliasing;

3. Coherent feature scales and domain sizes are implemented efficiently by cropping the Fourier spectrum, which offers a more comprehensive feature descriptor, at a low computational burden.

We integrate the L-SIFT descriptor into a DPM driven by SDM [11] and validate its performance in face alignment. We compare the performance of our Fourier domain designed filters with spatially-designed filters, and compare L-SIFT with conventional SIFT descriptors on popular face datasets. We further present the comparison against several state-of-the-art methods on two popular datasets: HELEN [133] and 300 Faces In-the-Wild (300-W) [134]. Experimental results show that L-SIFT as a part descriptor improves the performance of the DPM by a significant margin. The combined L-SIFT descriptor and SDM fitting algorithm achieves state-of-the-art performance on HELEN and 300-W common dataset, and comparable performance on the 300-W challenging dataset.

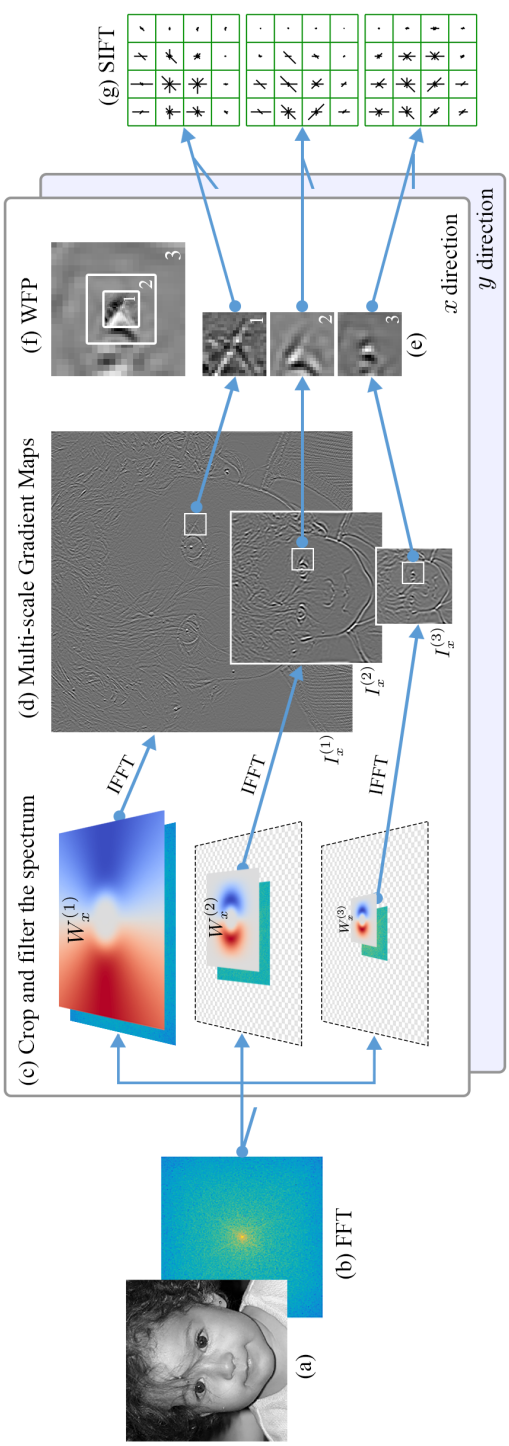
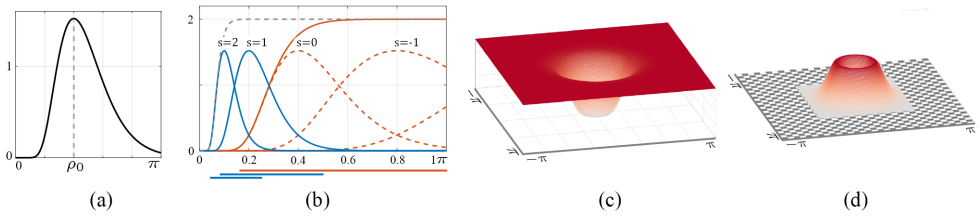


Figure A.1: Overview of extracting a loglet-SIFT part descriptor.



(a) (b) (c) (d)

Figure A.2: Filters in the Fourier domain. (a) A loglet function. (b) A loglet filterbank. Filters at higher resolution (red dashed) are accumulated to form the first scale filter (red solid). Additional lower scale filters are shown in blue. The x coordinate, which is the radius of the polar coordinate, becomes the scale dimension. The gray dashed-line indicates the summation of all the filters, which covers the pass-spectrum uniformly. The lines at the bottom show that each filter covers octaves of the lower frequency range. (c) The 2D high pass filter. (d) The first band pass filter. The checker-board area indicates the discarded frequencies.

A.1 Method

In this section we detail how to generate L-SIFT part descriptors for DPMs.

A.1.1 Feature scales

We start by decomposing an image into multiple channels with each preserving structures at certain scales. Describing the spectrum of an image in polar coordinates centred at the zero frequency, a frequency coordinate can be denoted by $\boldsymbol{u} = [\rho, \theta]$. The radius ρ actually represents a scale axis with larger scale (lower frequency) being closer to the origin. Therefore the scales can be decomposed and selected by arranging wavelets along the radius. We choose the loglets [30] as the basis functions.

The design of loglet functions for explicit scale selection has been introduced in Chapter 4. We review the functions in the scenario of feature decomposition. As has been defined in (4.1), a loglet function can be expressed by,

$$\mathcal{W}(\boldsymbol{u}; s) = \operatorname{erf}\left(\alpha \log\left(\beta^{s+\frac{1}{2}} \frac{\rho}{\rho_0}\right)\right) - \operatorname{erf}\left(\alpha \log\left(\beta^{s-\frac{1}{2}} \frac{\rho}{\rho_0}\right)\right) \tag{A.1}$$

which is a band pass filter, see Fig. A.2(a). erf is the error function and equals twice the integral of a normalised Gaussian function. α controls the radial bandwidth, s is an integer defining the scale of the filter, and $\beta > 1$ sets the relative ratio of adjacent scales – set to two for one octave intervals. ρ_0 is the peak radial frequency of the filter with scale $s = 0$.

To preserve sharp (small scale) textures of an image, the optimal filter should cover the higher frequency components. Note that a single filter is band pass, so we need to accumulate a group of filters successively having one-octave higher central frequencies,

$$\mathcal{W}^{(1)} = \sum_{s=0,-1,\ldots} \mathcal{W}(\boldsymbol{u}; s). \tag{A.2}$$

This achieves an even coverage towards the highest frequency benefiting from the uniformity property of loglets, see the red curve in Fig. A.2(b). The resultant 2D filter is shown in Fig. A.2(c). The filter accumulation enables a much larger radial bandwidth making it insensitive to scale changes. It is worth noting that the accumulation process is similar to the scale pooling used by Domain Size Pooling (DSP) [115], where local features across adjacent spatial scales are accumulated. The reason behind the better performance of DSP is that it marginalises the feature scales, which corresponds to a wider coverage of the frequency range. This is done in our approach explicitly with much lower computation burden. We prove the equivalence of *Fourier* filter accumulation and *spatial* scale pooling under certain approximations in Appendix E in the supplementary materials.

To obtain a more comprehensive description, we extract local features at additional larger spatial scales by using filters covering the complementary lower frequency range,

$$\mathcal{W}^{(s)}(\boldsymbol{u}) = \mathcal{W}(\boldsymbol{u}; s - 1). \tag{A.3}$$

Two adjacent larger scale filters at one octave intervals are shown in Fig. A.2(b) as blue curves.

As the image filtering can be implemented in the Fourier domain by multiplication, the filters can be efficiently applied in the standard way,

$$I^{(s)} = \mathcal{F}^{-1}(\mathcal{I} \cdot \mathcal{W}^{(s)}), \quad s = \{1, 2, ...\}, \tag{A.4}$$

in which $\mathcal{F}$ represents the Fourier Transform and $\mathcal{I}$ the spectrum of the image I. The image is thus decomposed into multiple channels $\{I^{(s)}\}$.

A.1.2 Domain sizes

Given a fiducial landmark, local patches can be extracted from the filtered image channels to obtain a multi-scale description. Larger scale textures should be described at coherently larger domain sizes and lower resolutions. We show that this is evident in the Fourier domain and can be achieved straightforwardly.

Note in Fig. A.2(b) that the two larger scale filters attenuate towards high frequency and the filter magnitude beyond $\pi/2$ and $\pi/4$ is almost zero, which means little or no frequency higher than these values is preserved in the subband channels. Therefore we can cut off these areas of the spectrum, which results in an efficient image downsampling without information loss or aliasing effect. With the cropping process, equation (A.4) becomes,

$$I^{(s)} = \mathcal{F}^{-1}(\mathcal{I}^{(s)} \cdot \mathcal{W}^{(s)}), s = \{1, 2, ...\}, \tag{A.5}$$

in which $\mathcal{I}^{(s)}$ is the cropped spectrum centred at the low frequency with $1/2^{(s-1)}$ size of the whole spectrum, $\mathcal{W}^{(s)}$ is the filter of same size as $\mathcal{I}^{(s)}$, see Fig. A.1(c). As a result, the resolution of the image channels is reduced by 2^s at scale s and a subband image pyramid is obtained, see an example in Fig. A.3. Note that the lowest frequency component is not covered in any of these channels as it represents the slowly varying, local mean-level containing mostly the illumination information.

At a given landmark, local patches of the same size are extracted from each

of the channels, giving a multi-scale feature description (Fig. A.1(e)). Although of same size in pixels, each patch represents twice the domain size and preserves one octave lower frequency components compared with its previous level. In this way a coherence between the domain size and the feature scale is achieved and the Wavelet Feature Pyramid (WFP) built (Fig. A.1(f)).

A.1.3 Orientations

The WFP built on multi-scale image channels can be applied to a number of intensity-based part descriptors in DPMs. Here we focus on integrating the scale selection property of loglets with the pooling power of SIFT descriptors. As SIFT performs a neighbourhood pooling on oriented gradients, we explain how to generate multi-scale gradient maps by further decomposing the non-oriented image channels into x and y components. The easiest way may seem to be by applying differential operators spatially on these channels. However the fact that differential filters take the form of an *imaginary anti-symmetrical* window in the Fourier domain (explained in Appendix F), we can neatly generate the oriented gradient maps directly by converting the loglets to imaginary odd-windows.

Specifically, imaginary sinusoidal functions at orthogonal orientations are added as directional parts, decomposing the spectrum into x and y components,

$$\begin{aligned} \mathcal{W}_x^{(s)}(\boldsymbol{u}) =& j\cos(\theta)\cdot\mathcal{W}^{(s)}(\boldsymbol{u}) \\ \mathcal{W}_y^{(s)}(\boldsymbol{u}) =& j\sin(\theta)\cdot\mathcal{W}^{(s)}(\boldsymbol{u}) \end{aligned} \tag{A.6}$$

where θ is the orientation of vector $\boldsymbol{u}$ and $j=\sqrt{-1}$. The oriented filters are shown in Fig. A.4(a). One problem which arises is that the high pass filter (scale one) in Fig. A.4(a) has larger magnitude around the Nyquist frequency (the margin of the Fourier spectrum), and its antisymmetrical shape gives $W_x(-\pi) = -W_x(\pi)$, therefore the spectrum is discontinuous across periods, which results in significant aliasing. For this reason most differential filters are designed to have zero magnitude

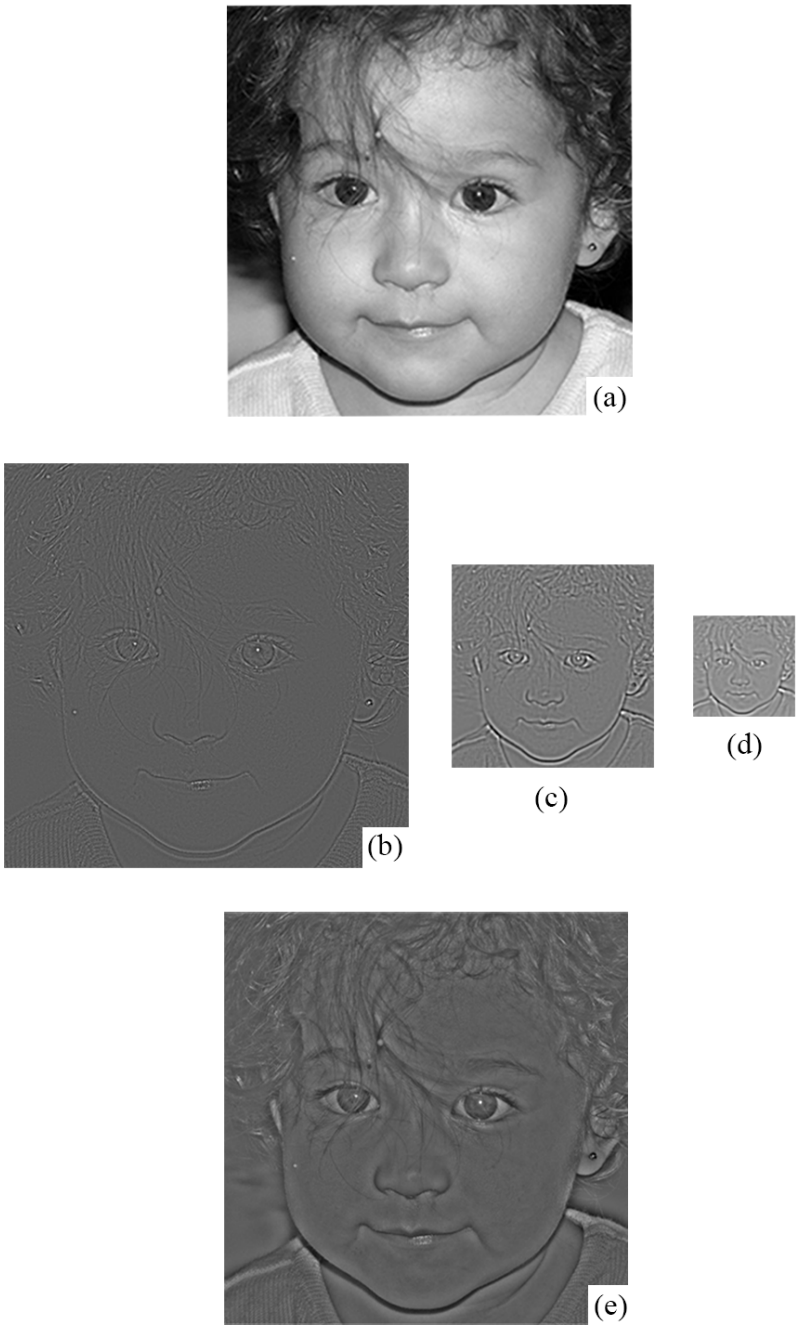


Figure A.3: (a) The original image. (b)(c)(d) Pyramid of multi-scale channels with increasing scales and reducing dimensions. (e) Summation of the three channels showing the image information captured. Note that the illumination (low varying components) is suppressed as the lowest frequency band of the spectrum is discarded.

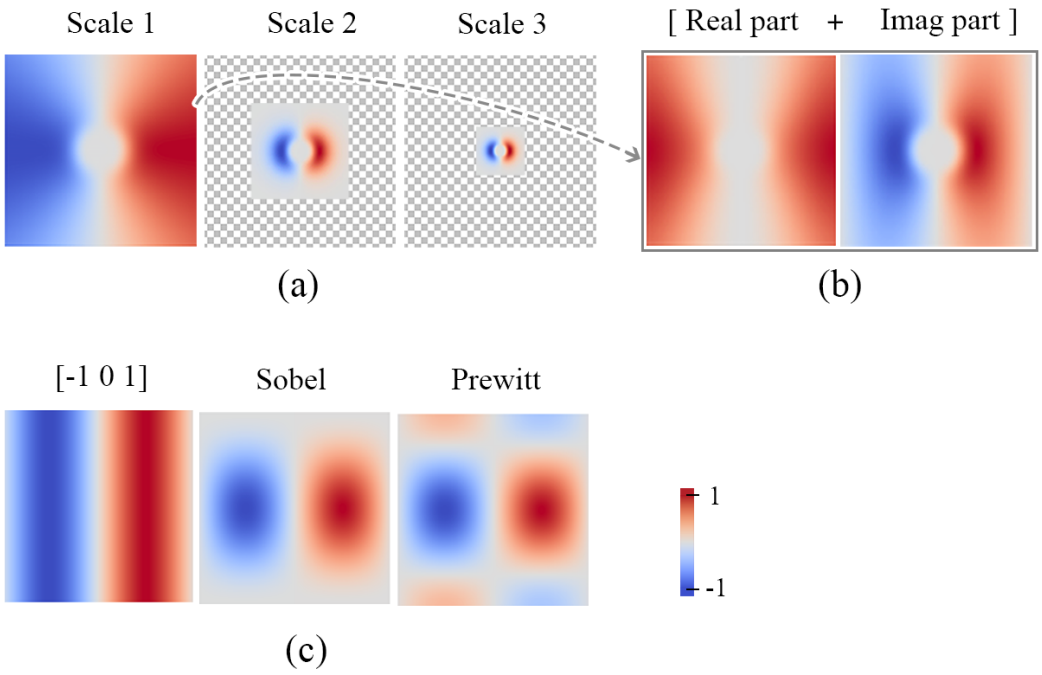


Figure A.4: Filters in the Fourier domain. (a) The imaginary parts of the oriented filter banks. The real parts are zero. (b) The real and imaginary part of the first scale filter after half-pixel shift. Note that the filter is now periodically continuous. (c) For comparison, spectra (imaginary parts) of spatially defined filters.

at the boundaries to prevent aliasing, but with the penalty of losing the highest frequency components thus sacrificing precision, see Fig. A.4(c). In our differential filtering, the highest frequency can be utilised without aliasing. The discontinuity is removed by adding a phase term, $e^{ju/2}$, to the odd filters,

$$\begin{aligned} \mathcal{W}_x(\boldsymbol{u}) =& e^{ju_x/2} \cdot j \cdot \cos(\theta) \cdot \mathcal{W}(\boldsymbol{u}) \\ \mathcal{W}_y(\boldsymbol{u}) =& e^{ju_y/2} \cdot j \cdot \sin(\theta) \cdot \mathcal{W}(\boldsymbol{u}) \end{aligned} \tag{A.7}$$

which results in a $\pi/2$ rotation in phase at one side $u_x = \pi$ and a $-\pi/2$ rotation at the other side $u_x = -\pi$, corresponding to a half-pixel shift in the spatial domain. The filters are now complex-valued and with continuity across periods, i.e., $\mathcal{W}_x(-\pi) = \mathcal{W}_x(\pi)$, see Fig. A.4(b).

The gradient map $\{I_x^{(s)}, I_y^{(s)}\}$ along x and y directions at multiple scales

can now be calculated by applying the oriented filters on the spectrum prior to the inverse Fourier Transform step. The L-SIFT descriptor is then obtained by calculating SIFTs on the resultant multi-scale gradient maps having equal block sizes in pixels. Because larger scale channels are down-sampled, the L-SIFT features represent real domain size and scales at octave intervals.

A.1.4 Loglet SIFT as part experts in DPM

We integrate the L-SIFT descriptor with the SDM algorithm [11] for facial landmark detection. Denote the L-SIFT descriptors at all landmarks as $h(I, \boldsymbol{s})$, with I being the image, $\boldsymbol{s}$ the landmarks, and $h(\cdot)$ the L-SIFT extracting function. In order to estimate the true landmark location $\boldsymbol{s}^*$ given an initial estimation $\hat{\boldsymbol{s}}$, we extract the descriptor $h(I, \hat{\boldsymbol{s}})$ at $\hat{\boldsymbol{s}}$ and learn the mapping $h(I, \hat{\boldsymbol{s}}) \rightarrow \Delta\boldsymbol{s}^*$, in which $\Delta\boldsymbol{s}^* = \boldsymbol{s}^* - \hat{\boldsymbol{s}}$. The direct mapping function satisfying all the cases in the dataset is non-linear in nature and can be over-fitted. So we adopt the SDM algorithm and approximate the non-linear mapping with a sequence of linear mapping $\{R^{(i)}, \boldsymbol{b}^{(i)}\}$ and landmark updating steps,

$$\begin{cases} \text{Mapping:} & \Delta\boldsymbol{s}^{(i)} = R^{(i)} h(I, \hat{\boldsymbol{s}}^{(i)}) + \boldsymbol{b}^{(i)}, \\ \text{Updating:} & \hat{\boldsymbol{s}}^{(i+1)} = \hat{\boldsymbol{s}}^{(i)} + \Delta\boldsymbol{s}^{(i)}. \end{cases} \tag{A.8}$$

The descriptor $h(I, \hat{\boldsymbol{s}}^{(i)})$ is extracted and updated at each iteration. Further details on SDM have been given in Section 4.2 and can also be found at [11].

A.2 Experiments

We show the performance of the L-SIFT descriptor on the problem of face alignment with DPM. We compare our filters with spatial domain gradient filters, evaluate the improvement brought to the DPM by the proposed L-SIFT descriptors, and report the performance against state-of-the-art methods. The evaluation metric used for all

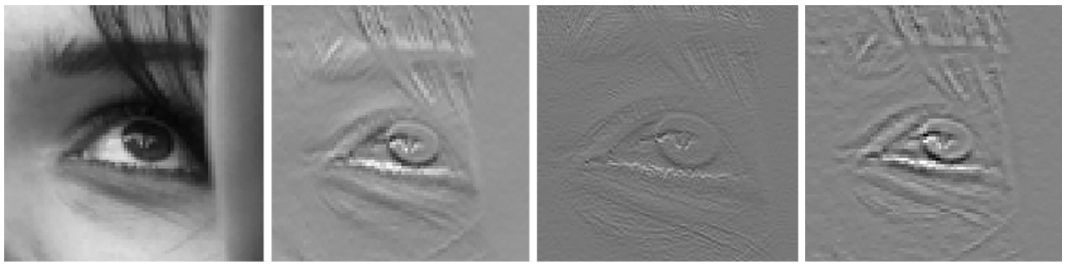
the face datasets is the root-mean-square (RMS) error normalised by the inter-pupil distance, as proposed in [135]. The parameters of the filter banks in all experiments are set as $\rho_0 = 0.3\pi$, $\alpha = 2$.

A.2.1 Datasets

The datasets for evaluation are the HELEN (194 landmarks), HELEN (68 landmarks), LFPW (68 landmarks) dataset and the 300-W dataset. The original HELEN (194 landmarks) are high-resolution facial images from the Flickr website. The dataset consists of 2000 training and 330 test images. The images are hand-annotated with 194 landmarks using Amazon Mechanical Turk to precisely locate the eyes, nose, mouth, eyebrows, and jawline. The HELEN (68 landmarks) dataset consists of facial images from original HELEN but re-annotated with 68 landmarks by the iBUG group [136]. The Labeled Face Parts in the Wild (LFPW) dataset [135] consists of 1432 faces from images downloaded from the web using simple text queries on sites such as google, flickr and yahoo. Each image is labelled by 29 fiducial points. The dataset is later on re-annotated by the iBUG group using 68 landmarks, denoted in this thesis by LFPW (68 landmarks). 300-W [137, 138, 139] is created from existing datasets including LFPW, AFW, HELEN, XM2VTS and the new iBUG dataset. Training and testing protocols can be found in the iBUG group web page [136].

A.2.2 Evaluation

Comparison with other differential filters. To demonstrate the contributions of the advanced gradient filters and the multi-scale features, we first compare the single scale gradient maps generated by our first scale filter $\mathcal{W}^{(1)}$ (Fig. A.4(a)) with conventional first order differential filters which can be used in SIFT descriptors, on the HELEN (68 landmarks) dataset. We show an example of a gradient map generated by these filters in Fig. A.5. We can see that the proposed filter better



(a) (b) (c) (d)

Figure A.5: Illustrative comparison of differential filters. Shown are the y direction gradients produced by: (b) $[-1, 0, 1]$, (c) $[-1, 1]$ and (d) our loglets bundle.

Table A.1: Comparison between SIFT built on spatial filters and our filters, on HELEN (68 landmarks) dataset

Filters	[-1 0 1]	[-1 1]	Sobel	Prewitt	$\mathcal{W}^{(1)}$	L-SIFT
Error	6.05	6.24	5.93	5.92	5.72	**5.21**

preserves sharper local structures. The SIFTs are calculated on these gradients and used as the part descriptors in SDM. The results are given in Table A.1. The result of the single scale filter $\mathcal{W}^{(1)}$ shows that simply replacing the conventional gradient map with the one by our filter improves the performance. We believe this benefits from the superior properties of the loglets over spatial-designed filters, as well as the larger bandwidth achieved by the filter accumulation. We further evaluate the performance of the proposed multi-scale L-SIFT descriptor with coherent feature scales and domain sizes. The result in Table. A.1 shows an additional significant improvement.

For efficiency purposes, the filter banks can be pre-calculated and stored. The most expensive computation for generating the feature is computing the gradient maps by applying filter banks in the Fourier domain. For the single level feature, there is no additional computation when compared to a conventional SIFT based on a spatial defined operators. For a feature pyramid with s levels, the computation includes a Fourier Transform, s element-wise matrix products and inverse Fourier

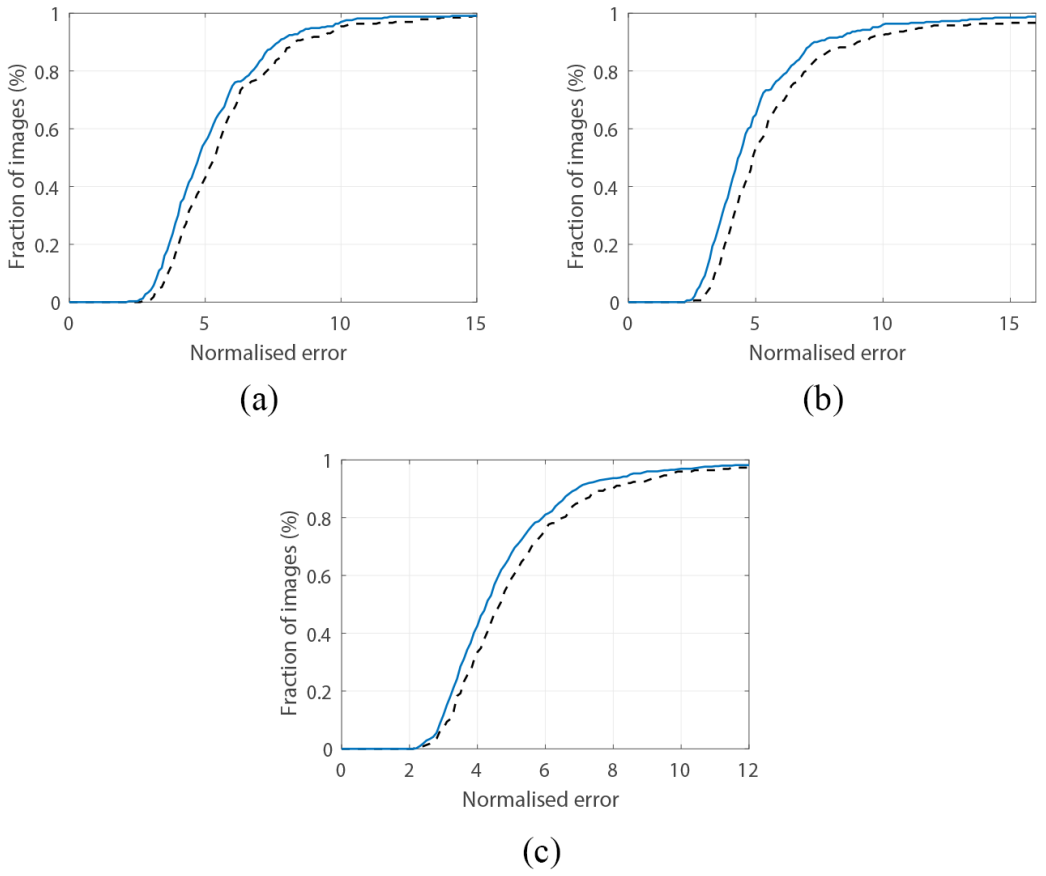


Figure A.6: Improvement brought to the SDM by L-SIFT on: (a) HELEN (194 landmarks), (b) HELEN (68 landmarks), (c) LFPW (68 landmarks). Deshed line: SDM with SIFT; Solid line: SDM with L-SIFT.

Transforms, both with reduced dimensions. This computation only need to be performed once before iteratively fitting the DPM to an image. Our MATLAB implementation for 3-scale features takes 9.7 ms on an image of size 400×400 using a 3.2GHz quad-core machine.

Improvement brought to the SDM. We evaluate the improvement brought to the SDM by the L-SIFT descriptor on several datasets including the HELEN [133] (194 landmarks), HELEN (68 landmarks) and LFPW (68 landmarks) dataset. The results are shown in Fig. A.6 and summarised in Table A.2. We can see an improvement brought to the SDM in all datasets.

Table A.2: Average error of landmark fitting.

	Helen (194)	Helen(68)	LFPW(68)
SDM(SIFT)	5.85	6.05	5.32
SDM(L-SIFT)	**5.30**	**5.21**	**4.90**
Improvement	9.4%	13.9%	7.7%

Table A.3: Average error of methods compared on HELEN (194 landmarks) dataset.

Method	RCPR[140]	ESR[129]	LBF fast[141]	LBF[141]	SDM(SIFT)[11]	**SDM(L-SIFT)**
Error	6.50	5.70	5.80	5.41	*5.85*	**5.30**

A.2.3 Comparison with the state-of-the-art face alignment

We compare our method with state-of-art benchmarks on the HELEN (194 landmarks) and 300-W datasets (68 landmarks) [139]. The testing set of 300-W is divided into a 'challenging' subset consisting of iBUG data and a 'common' subset consisting of the testing sets from HELEN and LFPW. The results are reported in table A.3 and A.4. For comparison with other methods, we list the original results in the literature.

Table A.4: Average error of methods compared on 300-W dataset.

Method	Common Subset	Challenging Subset
ESR[129]	5.28	17.00
LBF fast[141]	5.38	15.50
LBF[141]	4.95	**11.98**
SDM(SIFT) [11]	*5.60*	*15.40*
SDM(L-SIFT)	***4.91***	*13.49*

On the HELEN dataset, the improvement by the Fourier domain designed gradient filters is more significant and the combined SDM+L-SIFT algorithm outperforms the state-of-the-art methods. On the iBUG 300-W dataset, the combined

Figure A.7: Qualitative results from HELEN (top row) and 300-W challenging dataset (bottom row). The SDM with L-SIFT descriptors is compared against the one with SIFT. Green points show the ground truth, and the red points the fitting results.

algorithm gives best results in the common subset. Although it is not as precise in the challenging subset mainly due to the large pose variations of the faces, it still improves the performance of the SDM by a useful margin. We present qualitative results on particularly challenging cases in Fig. A.7 evaluating the improvement to the SDM algorithm. The results show that our feature descriptors yield better fitting performance especially on images with poor illumination or greater noise.

A.3 Summary

This chapter presents a part descriptor combining loglets and SIFT. The uniform coverage of the highest frequency gives no resolution loss and preserves the sharpest textures. Additional low frequency components are extracted, with coherently larger domain sizes achieved by cropping the Fourier spectrum, resulting in a more comprehensive feature description.

The combination of loglets and SIFT can be interpreted as an enhancement to a number of *invariances*, i.e, the invariance to illumination by the local pooling of SIFT and the suppression of slow varying mean level by the wavelets, as well as the invariances to noise by SIFT, and to sample shift by loglets. These properties improve the robustness of the descriptor to extrinsic variations. The proposed L-SIFT can be readily integrated in other Deformable Part Models for similar computer vision tasks and in medical image analysis.

APPENDIX B

The Closed-form Solution to the ML Shape

The maximum likelihood shape is the one minimising the energy function,

$$E(\boldsymbol{s}) = \frac{1}{2}\boldsymbol{b}^T\Lambda^{-1}\boldsymbol{b} + \frac{1}{2\rho}(||\boldsymbol{s}-\bar{\boldsymbol{s}}||^2 - ||\boldsymbol{b}||^2) + \sum_{l=1}^{L}\sum_{i=1}^{N}\frac{(\boldsymbol{x}_i - \hat{\boldsymbol{x}}_{i,l})^2}{2\boldsymbol{\sigma}_{i,l}^2}, \tag{B.1}$$

which can be rewritten in a matrix form,

$$E(\boldsymbol{s}) = \frac{1}{2}\boldsymbol{b}^T\Lambda^{-1}\boldsymbol{b} + \frac{1}{2\rho}\left((\boldsymbol{s}-\bar{\boldsymbol{s}})^T(\boldsymbol{s}-\bar{\boldsymbol{s}}) - \boldsymbol{b}^T\boldsymbol{b}\right) + \frac{1}{2}\sum_{l=1}^{L}(\boldsymbol{s}-\hat{\boldsymbol{s}}_l)^T\Sigma_l^{-1}(\boldsymbol{s}-\hat{\boldsymbol{s}}_l) \tag{B.2}$$

where $\Lambda = \text{diag}([\lambda_1, ..., \lambda_t])$ and $\Sigma_l = \text{diag}([\boldsymbol{\sigma}_{1,l}^2, ..., \boldsymbol{\sigma}_{N,l}^2])$, $\boldsymbol{b}$ is the vector of shape parameters and $\boldsymbol{s}$ is the shape. Equation B.2 has the typical form of an energy function for shape regularisation, with the notable difference that the second term is a summation of multiple predictions.

Note that $\boldsymbol{b}$ is calculated by,

$$\boldsymbol{b} = P^T(\boldsymbol{s}-\bar{\boldsymbol{s}}). \tag{B.3}$$

Substituting (B.3) into (B.2) gives,

$$\begin{aligned} E(\boldsymbol{s}) =& \frac{1}{2}(\boldsymbol{s}-\bar{\boldsymbol{s}})^T P\Lambda^{-1}P^T(\boldsymbol{s}-\bar{\boldsymbol{s}}) + \frac{1}{2\rho}\left((\boldsymbol{s}-\bar{\boldsymbol{s}})^T(\boldsymbol{s}-\bar{\boldsymbol{s}}) - (\boldsymbol{s}-\bar{\boldsymbol{s}})^T PP^T(\boldsymbol{s}-\bar{\boldsymbol{s}})\right) \\ &+ \frac{1}{2}\sum_{l=1}^{L}(\boldsymbol{s}-\hat{\boldsymbol{s}}_l)^T\Sigma_l^{-1}(\boldsymbol{s}-\hat{\boldsymbol{s}}_l). \end{aligned} \tag{B.4}$$

The optimal value of $\boldsymbol{s}$ is the one minimising $E(\boldsymbol{s})$, obtained by solving the equation:

$$\frac{\mathrm{d}E(\boldsymbol{s})}{\mathrm{d}\boldsymbol{s}} = P\Lambda^{-1}P^T(\boldsymbol{s}-\bar{\boldsymbol{s}}) + \frac{1}{\rho}(I-PP^T)(\boldsymbol{s}-\bar{\boldsymbol{s}}) + \sum_{l=1}^{L}\Sigma_l^{-1}(\boldsymbol{s}-\hat{\boldsymbol{s}}_l) = 0. \tag{B.5}$$

The solution is,

$$\boldsymbol{s} = A^{-1}B + \bar{s}, \tag{B.6}$$

in which,

$$\begin{aligned} A =& P\Lambda^{-1}P^T + \frac{1}{\rho}(I-PP^T) + \sum_{l=1}^{L}\Sigma_l^{-1}, \\ B =& \sum_{l=1}^{L}\Sigma_l^{-1}(\hat{\boldsymbol{s}}_l - \bar{\boldsymbol{s}}). \end{aligned} \tag{B.7}$$

APPENDIX C

Explicit Scale Selection in the Fourier Domain

The Fourier spectrum of an image spans a scale space, with lower frequencies representing larger scales spreading outwards from the centre point, see Fig. C.1. Selecting the scale ranges and decomposing the spectrum directly can introduce sharp boundaries therefore cause heavy aliasing, see an example in Fig. C.2.

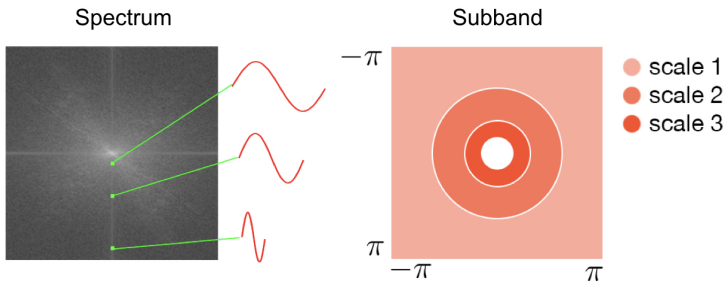


Figure C.1: Scale in a Fourier spectrum.

To avoid the aliasing we are looking for a basis function which has the following properties,

1. Smooth, or in wavelet vocabulary to have large vanishing moments;
2. Uniform which means the sum over scales is a constant and the windows cover the spectrum evenly.

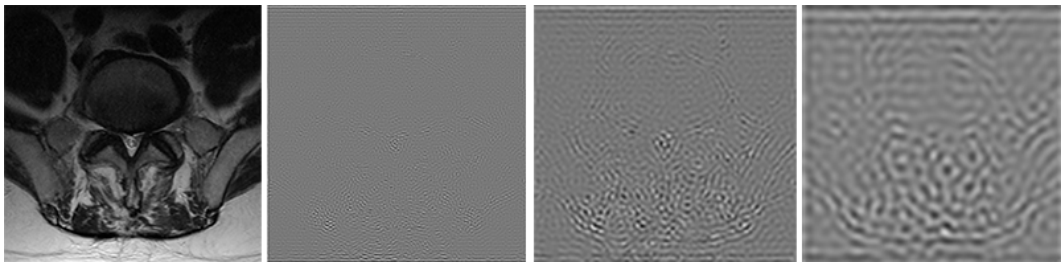


Figure C.2: Image aliasing caused by direct scale selection.

We choose the loglet functions for this purpose, and design the filter banks as shown in Fig. C.3. Now the scales can be decomposed and and the scale ranges can be selected without causing any aliasing, see Fig. C.4

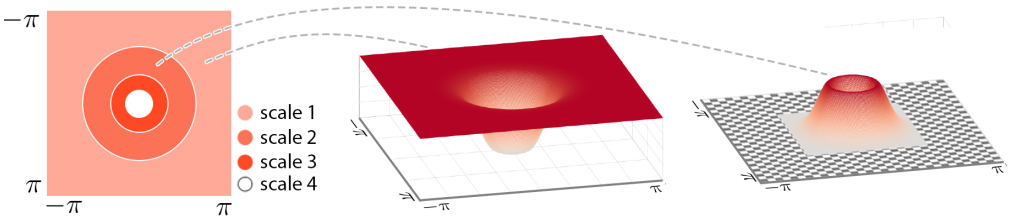


Figure C.3

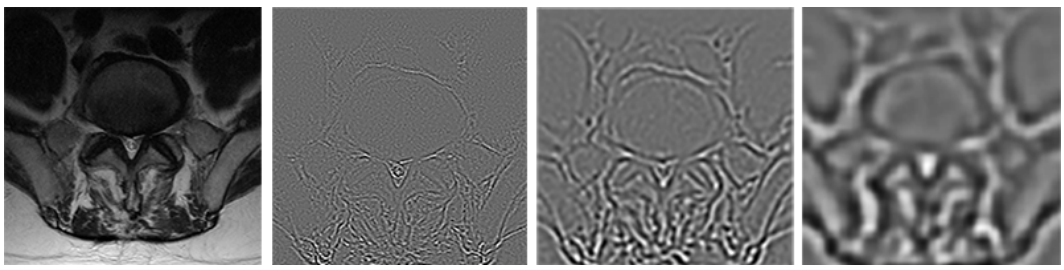


Figure C.4

APPENDIX D

Spectrum Cropping as Image Downsampling

A digital image is a discrete (i.e., band limited) sampling of the continuous (i.e., not band limited) *true* signal. In Fourier domain, the spectrum of the image therefore covers only the low frequency components of the true signal spectrum which spreads infinitely, and cuts off at the Nyquist boundary, see Fig.1. Images representing the same scene with lower resolution is presented in Fourier domain as a spectrum covering a smaller range centred at the zero frequency. As such, image downsampling can be done by cutting the spectrum keeping only the lower frequency. In practice, due to the discrete form, both the image and the spectrum are periodic signals. To avoid the aliasing caused by the periodic discontinuity, a window function such as Gaussian is applied to attenuate the magnitude near Nyquist frequency, which appears in spatial domain as a Gaussian smoothing. Below we describe the process mathematically.

Denote $\mathcal{I}$ as the spectrum of a digital image I, the image can be recovered by inverse Fourier transform,

$$I(\boldsymbol{x}) = \iint_{-\pi}^{\pi} \mathcal{I}(\boldsymbol{u}) e^{j\boldsymbol{x}\cdot\boldsymbol{u}} \mathrm{d}\boldsymbol{u} \tag{D.1}$$

where $\boldsymbol{x} = (x, y)$ and $\boldsymbol{u} = (u, v)$ are the index vectors in spatial and Fourier domain

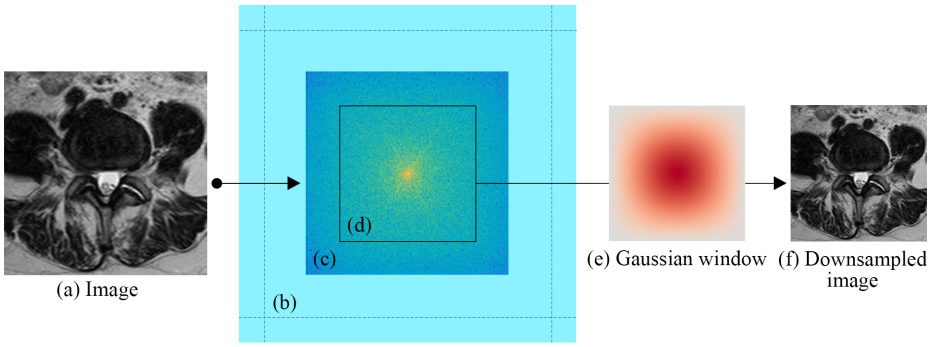


Figure D.1: (a) A digital image is a discrete sampling and approximation of the true continuous signal at certain resolution. (b) Spectrum of the true continuous signal is not band-limited therefore spreads infinitely. (c) Spectrum of the digital image is band-limited therefore covers only a low frequency area and is truncated at certain range. (d) Image downsampling can be performed directly by cropping the spectrum in Fourier domain. In theory the new spectrum (d) should be a subregion of the true spectrum (b), which is however not available. Therefore the cropping can be performed on the spectrum (c) instead. In practice cropping the spectrum can not ensure the periodic contineous of the new spectrum therefore causes alising effect. A standard way is to attenuate the spectrum to zero at the boundary by applying a Gaussian window to it (e), which corresponds to a smoothing in the spatial domain.

respectively. Suppose we want to downsample the image at ratio $\beta \in (0, 1)$. We first apply a window function to attenuate the components beyond the boundary $\pm\beta\pi$ to zero. With the windowing (D.1) equals to,

$$I(\boldsymbol{x}) = \iint_{-\beta\pi}^{\beta\pi} \mathcal{I}(\boldsymbol{u}) e^{j\boldsymbol{x}\cdot\boldsymbol{u}} \mathrm{d}\boldsymbol{u} \tag{D.2}$$

We define new variables $\boldsymbol{u}_1 = \boldsymbol{u}/\beta$, $\boldsymbol{x}_1 = \beta\boldsymbol{x}$, and a coordinate transform of the spectrum $\mathcal{I}_1(\boldsymbol{u}_1) = \mathcal{I}(\boldsymbol{u})$. Substituting them in to (D.2) we have,

$$\begin{aligned} I(\boldsymbol{x}) =& \iint_{-\beta\pi}^{\beta\pi} \mathcal{I}_1(\boldsymbol{u}_1) e^{j\boldsymbol{x}_1\cdot\boldsymbol{u}_1} \mathrm{d}(\beta\boldsymbol{u}_1) \\ =& \beta \iint_{-\pi}^{\pi} \mathcal{I}_1(\boldsymbol{u}_1) e^{j\boldsymbol{x}_1\cdot\boldsymbol{u}_1} \mathrm{d}\boldsymbol{u}_1 \\ =& \beta I_1(\boldsymbol{x}_1), \end{aligned} \tag{D.3}$$

where I_1 is the downsampled image, i.e.,

$$I_1(\boldsymbol{x}_1) = I(\boldsymbol{x})/\beta, \quad \boldsymbol{x}_1 = \beta\boldsymbol{x}. \tag{D.4}$$

In our case, with larger scale filters $\mathcal{W}^{(s)}, s \in \{2, 3, ...\}$, the high frequency components beyond the boundary $\pm\pi/2^{(s-1)}$ are eliminated and a direct crop of the spectrum gives an efficient downsampling at ratio $1/2^{(s-1)}$ without aliasing.

APPENDIX E

Scale Pooling in Spatial Domain and Filter Accumulation in Fourier Domain

One of the key features of the SIFT is that it performs the pooling in a 2D neighbourhood in the spatial domain. In a recent study [115] Dong proposed to extend the pooling to spatial-scale space, by performing an additional pooling across adjacent scales in order to enhance the invariance to minor scale changes. We prove that the filter accumulation in Fourier domain is equivalent to scale pooling, under the approximation that gradients at adjacent scales have similar orientations, which is reasonable when low orientation resolution such as $\pi/4$ (8 orientation bins) are used.

An un-normalised gradient histogram of a SIFT cell in a region centred at point $\boldsymbol{x}$ can be written compactly as [115],

$$h(\theta|I) = \sum_{\boldsymbol{x}'} \kappa_\epsilon(\theta - \angle\nabla I(\boldsymbol{x}'))\kappa_\sigma(\boldsymbol{x} - \boldsymbol{x}')||\nabla I(\boldsymbol{x}')|| \tag{E.1}$$

where θ is a variable corresponding to an orientation histogram bin. Discrete bins are computed using a bilinear interpolation kernel κ_ϵ with $\epsilon = 2\pi/n$ where n is the number of bins, and linear spatial weighting kernel κ_σ with σ controls the size of a

cell. Note that unlike in [115] we use a discrete form in (E.1) as is the case in the practical implementation.

Now consider the filter accumulation in Fourier domain used in the first scale of our L-SIFT descriptor. The gradient can be represented by $\nabla I = [I_x, I_y]$, with each direction calculated with the first scale filter (a bundle of loglets),

$$\begin{aligned} I_x =& \mathcal{F}^{-1}(\mathcal{F}(I) \cdot \mathcal{W}_x^{(1)}) \\ =& \mathcal{F}^{-1}\left(\mathcal{F}(I) \cdot \sum_s \mathcal{W}_x(\boldsymbol{u}, s)\right) \\ =& \sum_s \mathcal{F}^{-1}\left(\mathcal{F}(I) \cdot \mathcal{W}_x(\boldsymbol{u}, s)\right), \quad s \in \{0, -1, ...\} \end{aligned} \tag{E.2}$$

The gradient can therefore be written as

$$\nabla I = [I_x, I_y] = \sum_s \nabla^{(s)} I \tag{E.3}$$

with $\nabla^{(s)}$ represents a gradient computation with a single loglet filter at scale s. Substituting (E.3) into (E.1) we obtain the SIFT with gradient computed by accumulated filters,

$$h(\theta|I) = \sum_{\boldsymbol{x}'} \kappa_\epsilon(\theta - \angle \nabla I(\boldsymbol{x}'))\kappa_\sigma(\boldsymbol{x} - \boldsymbol{x}')|| \sum_s \nabla^{(s)} I(\boldsymbol{x}')|| \tag{E.4}$$

Considering that feature gradients at adjacent scales have similar orientations, and SIFT uses discrete orientation bins with a low angular resolution, the following approximation can be made,

$$\begin{aligned} h(\theta|I) \approx& \sum_s \sum_{\boldsymbol{x}'} \kappa_\epsilon(\theta - \angle \nabla_s I(\boldsymbol{x}'))\kappa_\sigma(\boldsymbol{x} - \boldsymbol{x}')||\nabla_s I(\boldsymbol{x}')|| \\ =& \sum_s h_s(\theta|I) \end{aligned} \tag{E.5}$$

Where $h_s(\theta|I)$ represents a standard SIFT with the gradient computed by a single

loglet at scale s. Therefore we proved that *a SIFT computed on the gradient by accumulated filters in Fourier domain is equivalent to accumulating a group of SIFTs cumputed on multi-scale gradients in spatial domain.*

The rationale behind the significant improvement by scale pooling is that it gives invariance to minor scale changes as well as sample shifts of digital images. In our strategy, the first invariance is achieved by expanding the bandwidth by filter accumulation, the second invariance comes from the natural insensitivity of loglets function to sample shift [30].

APPENDIX F

Derivative of Images in Fourier Domain

Theorem: Derivative of an image $I(x, y)$ results in an imaginary anti-symmetrical transform of the spectrum $\mathcal{F}(I)$.

Derivation: Taking the derivative with respect to x as an example. The Fourier transform of I is,

$$\mathcal{F}(I) = \iint_{-\infty}^{+\infty} I \; e^{-2\pi i(ux+vy)} dxdy, \tag{F.1}$$

where $(u, v) \in [-\pi, \pi]$ are the coordinates in the Fourier domain. The Fourier transform of the derivative $\mathrm{d}I/\mathrm{d}x$ is,

$$\mathcal{F}(\frac{\mathrm{d}I}{\mathrm{d}x}) = \iint_{-\infty}^{+\infty} \frac{\mathrm{d}I}{\mathrm{d}x} e^{-2\pi i(ux+vy)} dxdy, \tag{F.2}$$

Consider the following equation,

$$\begin{aligned} \frac{\mathrm{d}(Ie^{-2\pi i(ux+vy)})}{\mathrm{d}x} =& \frac{\mathrm{d}I}{\mathrm{d}x} e^{-2\pi i(ux+vy)} + \frac{\mathrm{d}(e^{-2\pi i(ux+vy)})}{\mathrm{d}x} I \\ =& \frac{\mathrm{d}I}{\mathrm{d}x} e^{-2\pi i(ux+vy)} - 2\pi iu \; e^{-2\pi i(ux+vy)} I, \end{aligned} \tag{F.3}$$

therefore,

$$\frac{\mathrm{d}I}{\mathrm{d}x} e^{-2\pi i(ux+vy)} = \frac{\mathrm{d}(Ie^{-2\pi i(ux+vy)})}{\mathrm{d}x} + 2\pi iu \; e^{-2\pi i(ux+vy)} I, \tag{F.4}$$

Substituting (F.4) into (F.2) we have,

$$\begin{aligned}\mathcal{F}(\frac{\mathrm{d}I}{\mathrm{d}x}) =& \iint_{-\infty}^{+\infty} \frac{\mathrm{d}(Ie^{-2\pi i(ux+vy)})}{\mathrm{d}x} dxdy + \iint_{-\infty}^{+\infty} 2\pi iu\ e^{-2\pi i(ux+vy)} I dxdy \\ =& 2\pi iu \iint_{-\infty}^{+\infty} I\ e^{-2\pi i(ux+vy)} dxdy \\ =& 2\pi iu\ \mathcal{F}(I),\end{aligned} \tag{F.5}$$

which indicates that the derivative of image I results in the multiplication of the Fourier spectrum by an imaginary anti-symmetrical term $2\pi iu$.

Bibliography

[1] Doi, K.: Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Computerized medical imaging and graphics **31**(4) (2007) 198–211

[2] Beutel, J., Kundel, H.L., Van Metter, R.L.: Handbook of medical imaging: Physics and psychophysics. Volume 1. Spie Press (2000)

[3] Shen, D., Wu, G., Zhang, D., Suzuki, K., Wang, F., Yan, P.: Machine learning in medical imaging. Comp. Med. Imag. and Graph. **41** (2015) 1–2

[4] Deyo, R.A.: Treatment of lumbar spinal stenosis: a balancing act. The Spine Journal **10**(7) (2010) 625–627

[5] Watters, W.C., Baisden, J., Gilbert, T.J., Kreiner, S., Resnick, D.K., Bono, C.M., Ghiselli, G., Heggeness, M.H., Mazanec, D.J., O'Neill, C., et al.: Degenerative lumbar spinal stenosis: an evidence-based clinical guideline for the diagnosis and treatment of degenerative lumbar spinal stenosis. The spine journal **8**(2) (2008) 305–310

[6] Miettinen, O.S., Flegel, K.M.: Elementary concepts of medicine. Journal of evaluation in clinical practice **9**(3) (2003) 315–317

[7] Genevay, S., Atlas, S.J., Katz, J.N.: Variation in eligibility criteria from studies of radiculopathy due to a herniated disc and of neurogenic claudication

due to lumbar spinal stenosis: a structured literature review. Spine **35**(7) (2010) 803

[8] Steurer, J., Roner, S., Gnannt, R., Hodler, J.: Quantitative radiologic criteria for the diagnosis of lumbar spinal stenosis: a systematic literature review. BMC musculoskeletal disorders **12**(1) (2011) 175

[9] Ericksen, S.: Lumbar spinal stenosis: Imaging and non-operative management. In: Seminars in Spine Surgery. Volume 25., Elsevier (2013) 234–245

[10] Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(6) (2001) 681–685

[11] Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE (2013) 532–539

[12] Lindner, C., Thiagarajah, S., Wilkinson, J., Consortium, T., Wallis, G., Cootes, T.F.: Fully automatic segmentation of the proximal femur using random forest regression voting. Medical Imaging, IEEE Transactions on **32**(8) (2013) 1462–1472

[13] Antonakos, E., Alabort-i Medina, J., Zafeiriou, S.: Active Pictorial Structures. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 5435–5444

[14] Saragih, J.M., Lucey, S., Cohn, J.F.: Face alignment through subspace constrained mean-shifts. In: IEEE 12th International Conference on Computer Vision, IEEE (2009) 1034–1041

[15] Cristinacce, D., Cootes, T.F.: Boosted Regression Active Shape Models. In: BMVC. (2007) 1–10

[16] Baker, S., Matthews, I.: Equivalence and efficiency of image alignment algorithms. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Volume 1., IEEE (2001) I–1090

[17] Simonyan, K., Vedaldi, A., Zisserman, A.: Descriptor learning using convex optimisation. In: European Conference on Computer Vision, Springer (2012) 243–256

[18] Simonyan, K., Vedaldi, A., Zisserman, A.: Learning local feature descriptors using convex optimisation. IEEE Transactions on PAMI **36**(8) (2014) 1573–1585

[19] Xiao, L.: Dual Averaging Method for Regularized Stochastic Learning and Online Optimization. In: Advances in Neural Information Processing Systems. Curran Associates, Inc. (2009) 2116–2124

[20] Xiao, L.: Dual averaging methods for regularized stochastic learning and online optimization. Journal of Machine Learning Research **11**(Oct) (2010) 2543–2596

[21] Chen, Y., Bi, J., Wang, J.Z.: Miles: Multiple-instance learning via embedded instance selection. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(12) (2006) 1931–1947

[22] Lootus, M., Kadir, T., Zisserman, A.: Vertebrae detection and labelling in lumbar MR images. In: MICCAI CSI Workshop. Springer (2013) 219–230

[23] Zhao, Q., Okada, K., Rosenbaum, K., Kehoe, L., Zand, D.J., Sze, R., Summar, M., Linguraru, M.G.: Digital facial dysmorphology for genetic screening: Hierarchical constrained local model using ICA. Medical Image Analysis **18**(5) (2014) 699–710

[24] Shen, W., Zhou, M., Yang, F., Yang, C., Tian, J.: Multi-scale convolutional neural networks for lung nodule classification. In: International Conference on Information Processing in Medical Imaging, Springer (2015) 588–599

[25] Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M.: Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging **35**(5) (2016) 1285–1298

[26] Schlegl, T., Waldstein, S.M., Vogl, W.D., Schmidt-Erfurth, U., Langs, G.: Predicting semantic descriptions from medical images with convolutional neural networks. In: International Conference on IPMI, Springer (2015) 437–448

[27] Mahapatra, D.: Retinal image quality classification using saliency maps and cnns. In: International Conference on MICCAI, Springer (2016)

[28] Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free?-weakly-supervised learning with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 685–694

[29] Jamaludin, A., Kadir, T., Zisserman, A.: SpineNet: Automatically Pinpointing Classification Evidence in Spinal MRIs. In: International Conference on MICCAI, Springer (2016) 166–175

[30] Knutsson, H., Andersson, M.: Loglets: Generalized quadrature and phase for local spatio-temporal structure estimation. In: Image Analysis. Springer (2003) 741–748

[31] Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. Computer Vision and Image Understanding **61**(1) (1995) 38–59

[32] Davatzikos, C., Tao, X., Shen, D.: Hierarchical active shape models, using the wavelet transform. IEEE transactions on medical imaging **22**(3) (2003) 414–423

[33] Yan, P., Xu, S., Turkbey, B., Kruecker, J.: Discrete deformable model guided by partial active shape model for TRUS image segmentation. IEEE Transactions on Biomedical Engineering **57**(5) (2010) 1158–1166

[34] Spiegel, M., Hahn, D.A., Daum, V., Wasza, J., Hornegger, J.: Segmentation of kidneys using a new active shape model generation technique based on non-rigid image registration. Computerized Medical Imaging and Graphics **33**(1) (2009) 29–39

[35] Zhang, Q., Bhalerao, A., Helm, E., Hutchinson, C.: Active shape model unleashed with multi-scale local appearance. In: Image Processing (ICIP), 2015 IEEE International Conference on, IEEE (2015) 4664–4668

[36] Hontani, H., Tsunekawa, Y., Sawada, Y.: Accurate and robust registration of nonrigid surface using hierarchical statistical shape model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2013) 2977–2984

[37] Chen, X., Udupa, J.K., Alavi, A., Torigian, D.A.: Gc-asm: Synergistic integration of graph-cut and active shape model strategies for medical image segmentation. Computer Vision and Image Understanding **117**(5) (2013) 513–524

[38] Parsons, C., Hutchinson, C., Helm, E., Clarke, A., Mirza, A.B., Zhang, Q., Bhalerao, A.: Development of an automated shape and textural software model of the paediatric knee for estimation of skeletal age.

[39] Jones, M.J., Poggio, T.: Multidimensional morphable models: A framework for representing and matching object classes. International Journal of Computer Vision **29**(2) (1998) 107–131

[40] Rueckert, D., Frangi, A.F., Schnabel, J.A.: Automatic construction of 3-d statistical deformation models of the brain using nonrigid registration. IEEE transactions on medical imaging **22**(8) (2003) 1014–1025

[41] Matthews, I., Baker, S.: Active appearance models revisited. International Journal of Computer Vision **60**(2) (2004) 135–164

[42] Antonakos, E., Alabort-i Medina, J., Tzimiropoulos, G., Zafeiriou, S.: Hog active appearance models. In: Image Processing (ICIP), 2014 IEEE International Conference on, IEEE (2014) 224–228

[43] Navarathna, R., Sridharan, S., Lucey, S.: Fourier active appearance models. In: 2011 IEEE International Conference on Computer Vision, IEEE (2011) 1919–1926

[44] Roberts, M., Cootes, T.F., Adams, J.E.: Vertebral morphometry: semiautomatic determination of detailed shape from dual-energy X-ray absorptiometry images using active appearance models. Investigative radiology **41**(12) (2006) 849–859

[45] Roberts, M.G.: Automatic detection and classification of vertebral fracture using statistical models of appearance. PhD thesis, University of Manchester (2008)

[46] Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. International Journal of Computer Vision **61**(1) (2005) 55–79

[47] Felzenszwalb, P.F., Girshick, R.B., McAllester, D.: Cascade object detection with deformable part models. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2010) 2241–2248

[48] Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(9) (2010) 1627–1645

[49] Cristinacce, D., Cootes, T.F.: Feature Detection and Tracking with Constrained Local Models. In: Proceedings of the British machine Vision Conference. Volume 1., Citeseer (2006) 3

[50] Cristinacce, D., Cootes, T.: Automatic feature localisation with constrained local models. Pattern Recognition **41**(10) (2008) 3054–3067

[51] Saragih, J.M., Lucey, S., Cohn, J.F.: Deformable model fitting by regularized landmark mean-shift. International Journal of Computer Vision **91**(2) (2011) 200–215

[52] Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2012) 2879–2886

[53] Brunet, N., Perez, F., De La Torre, F.: Learning good features for active shape models. In: Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on, IEEE (2009) 206–211

[54] Roberts, M.G., Cootes, T.F., Pacheco, E., Oh, T., Adams, J.E.: Segmentation of lumbar vertebrae using part-based graphs and active appearance models. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009. Springer (2009) 1017–1024

[55] Tresadern, P.A., Bhaskar, H., Adeshina, S.A., Taylor, C.J., Cootes, T.F.: Combining Local and Global Shape Models for Deformable Object Matching. In: Proceedings of the British Machine Vision Conference. Volume 9. (2009) 451–458

[56] Gee, J.C., Reivich, M., Bajcsy, R.: Elastically deforming 3d atlas to match anatomical brain images. Journal of computer assisted tomography **17**(2) (1993) 225–236

[57] Kalinić, H.: Atlas-based image segmentation: A survey. (2009)

[58] Stegmann, M.B.: Generative interpretation of medical images. Lingby: Thesis, University of Denmark (2004)

[59] Mitchell, S.C., Bosch, J.G., Lelieveldt, B.P., van der Geest, R.J., Reiber, J.H., Sonka, M.: 3-D active appearance models: segmentation of cardiac MR and ultrasound images. IEEE Transactions on Medical Imaging **21**(9) (2002) 1167–1178

[60] Andreopoulos, A., Tsotsos, J.K.: Efficient and generalizable statistical models of shape and appearance for analysis of cardiac MRI. Medical Image Analysis **12**(3) (2008) 335–357

[61] Bhatia, K.K., Hajnal, J.V., Puri, B.K., Edwards, A.D., Rueckert, D.: Consistent groupwise non-rigid registration for atlas construction. In: Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on, IEEE (2004) 908–911

[62] Lorenzen, P., Prastawa, M., Davis, B., Gerig, G., Bullitt, E., Joshi, S.: Multimodal image set registration and atlas formation. Medical image analysis **10**(3) (2006) 440–451

[63] Fillard, P., Pennec, X., Thompson, P., Ayache, N.: Evaluating brain anatomical correlations via canonical correlation analysis of sulcal lines. PhD thesis, INRIA (2007)

[64] Kirschner, M., Becker, M., Wesarg, S.: 3D active shape model segmentation with nonlinear shape priors. In: Medical Image Computing and Computer-Assisted Intervention. Springer (2011) 492–499

[65] Jafari-Khouzani, K., Soltanian-Zadeh, H.: Multiwavelet grading of pathological images of prostate. IEEE Transactions on Biomedical Engineering **50**(6) (2003) 697–704

[66] Salamanca, L., Vlassis, N., Diederich, N., Bernard, F., Skupin, A.: Improved parkinsons disease classification from diffusion mri data by fisher vector descriptors. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2015) 119–126

[67] Otálora, S., Cruz-Roa, A., Arevalo, J., Atzori, M., Madabhushi, A., Judkins, A.R., González, F., Müller, H., Depeursinge, A.: Combining unsupervised feature learning and riesz wavelets for histopathology image representation: application to identifying anaplastic medulloblastoma. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2015) 581–588

[68] Basavanhally, A., Ganesan, S., Feldman, M., Shih, N., Mies, C., Tomaszewski, J., Madabhushi, A.: Multi-field-of-view framework for distinguishing tumor grade in er+ breast cancer from entire histopathology slides. IEEE transactions on biomedical engineering **60**(8) (2013) 2089–2099

[69] Chen, H., Dou, Q., Ni, D., Cheng, J.Z., Qin, J., Li, S., Heng, P.A.: Automatic fetal ultrasound standard plane detection using knowledge transferred

recurrent neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2015) 507–514

[70] Song, X., Meng, L., Shi, Q., Lu, H.: Learning tensor-based features for whole-brain fmri classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2015) 613–620

[71] BenTaieb, A., Li-Chang, H., Huntsman, D., Hamarneh, G.: Automatic diagnosis of ovarian carcinomas via sparse multiresolution tissue representation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2015) 629–636

[72] Wong, K.C., Tee, M., Chen, M., Bluemke, D.A., Summers, R.M., Yao, J.: Computer-aided infarction identification from cardiac ct images: a biomechanical approach with svm. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2015) 144–151

[73] Chandran, V., Zysset, P., Reyes, M.: Prediction of trabecular bone anisotropy from quantitative computed tomography using supervised learning and a novel morphometric feature descriptor. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2015) 621–628

[74] Li, W., Zhang, J., McKenna, S.J.: Multiple instance cancer detection by boosting regularised trees. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2015) 645–652

[75] Wolterink, J.M., Leiner, T., Viergever, M.A., Išgum, I.: Automatic coronary calcium scoring in cardiac ct angiography using convolutional neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2015) 589–596

[76] Lian, C., Ruan, S., Denœux, T., Li, H., Vera, P.: Dempster-shafer theory based feature selection with sparse constraint for outcome prediction in cancer therapy. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2015) 695–702

[77] Liu, M., Lu, L., Ye, X., Yu, S., Salganicoff, M.: Sparse classification for computer aided diagnosis using learned dictionaries. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2011) 41–48

[78] Varol, E., Sotiras, A., Davatzikos, C.: Disentangling disease heterogeneity with max-margin multiple hyperplane classifier. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2015) 702–709

[79] Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10) (2015) 1993–2024

[80] Pan, Y., Huang, W., Lin, Z., Zhu, W., Zhou, J., Wong, J., Ding, Z.: Brain tumor grading based on neural networks and convolutional neural networks. In: Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE, IEEE (2015) 699–702

[81] Carneiro, G., Nascimento, J., Bradley, A.P.: Unregistered multiview mammogram analysis with pre-trained deep learning models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2015) 652–660

[82] van Ginneken, B., Setio, A.A., Jacobs, C., Ciompi, F.: Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed

tomography scans. In: Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on, IEEE (2015) 286–289

[83] Bar, Y., Diamant, I., Wolf, L., Lieberman, S., Konen, E., Greenspan, H.: Chest pathology detection using deep learning with non-medical training. In: Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on, IEEE (2015) 294–297

[84] Ciompi, F., de Hoop, B., van Riel, S.J., Chung, K., Scholten, E.T., Oudkerk, M., de Jong, P.A., Prokop, M., van Ginneken, B.: Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2d views and a convolutional neural network out-of-the-box. Medical image analysis **26**(1) (2015) 195–202

[85] Hofmanninger, J., Langs, G.: Mapping visual features to semantic profiles for retrieval in medical imaging. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 457–465

[86] Carneiro, G., Nascimento, J.C.: Combining multiple dynamic models and deep learning architectures for tracking the left ventricle endocardium in ultrasound data. IEEE transactions on pattern analysis and machine intelligence **35**(11) (2013) 2592–2607

[87] Li, R., Zhang, W., Suk, H.I., Wang, L., Li, J., Shen, D., Ji, S.: Deep learning based imaging data completion for improved brain disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2014) 305–312

[88] Zhang, Q., Bhalerao, A., Hutchinson, C.: Weakly-supervised evidence pinpointing and description. In: International Conference on Information Processing in Medical Imaging, Springer (2017)

[89] Xu, Y., Zhu, J.Y., Eric, I., Chang, C., Lai, M., Tu, Z.: Weakly supervised histopathology cancer image segmentation and classification. Medical image analysis **18**(3) (2014) 591–604

[90] Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence **89**(1) (1997) 31–71

[91] Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)

[92] Simoncelli, E.P., Freeman, W.T., Adelson, E.H., Heeger, D.J.: Shiftable multiscale transforms. Information Theory, IEEE Transactions on **38**(2) (1992) 587–607

[93] Gross, M.H., Koch, R.: Visualization of multidimensional shape and texture features in laser range data using complex-valued Gabor wavelets. Visualization and Computer Graphics, IEEE Transactions on **1**(1) (1995) 44–59

[94] Nestares, O., Navarro, R., Portilla, J., Tabernero, A.: Efficient spatial-domain implementation of a multiscale image representation based on Gabor functions. Journal of Electronic Imaging **7**(1) (1998) 166–173

[95] Ro, Y.M., Kim, M., Kang, H.K., Manjunath, B., Kim, J.: MPEG-7 homogeneous texture descriptor. ETRI journal **23**(2) (2001) 41–51

[96] Field, D.J.: Relations between the statistics of natural images and the response properties of cortical cells. JOSA A **4**(12) (1987) 2379–2394

[97] Fischer, S., Šroubek, F., Perrinet, L., Redondo, R., Cristóbal, G.: Self-invertible 2D log-Gabor wavelets. International Journal of Computer Vision **75**(2) (2007) 231–246

[98] Starck, J.L., Candès, E.J., Donoho, D.L.: The curvelet transform for image denoising. Image Processing, IEEE Transactions on **11**(6) (2002) 670–684

[99] Do, M.N., Vetterli, M.: The contourlet transform: an efficient directional multiresolution image representation. Image Processing, IEEE Transactions on **14**(12) (2005) 2091–2106

[100] Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics (2001) 1189–1232

[101] Saragih, J., Goecke, R.: A nonlinear discriminative approach to aam fitting. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, IEEE (2007) 1–8

[102] Tresadern, P.A., Sauer, P., Cootes, T.F.: Additive update predictors in active appearance models. In: BMVC. Volume 2., Citeseer (2010) 4

[103] Dollár, P., Welinder, P., Perona, P.: Cascaded pose regression. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE (2010) 1078–1085

[104] Rivera, S., Martinez, A.M.: Learning deformable shape manifolds. Pattern Recognition **45**(4) (2012) 1792–1801

[105] Cootes, T.F., Ionita, M.C., Lindner, C., Sauer, P.: Robust and accurate shape model fitting using random forest regression voting. In: European Conference on Computer Vision, Springer (2012) 278–291

[106] Sánchez-Lozano, E., De la Torre, F., González-Jiménez, D.: Continuous regression for non-rigid image alignment. In: European Conference on Computer Vision, Springer (2012) 250–263

[107] Zimmermann, K., Matas, J., Svoboda, T.: Tracking by an optimal sequence of linear predictors. IEEE transactions on pattern analysis and machine intelligence **31**(4) (2009) 677–692

[108] Xiong, X., De la Torre, F.: Supervised descent method for solving nonlinear least squares problems in computer vision. arXiv preprint arXiv:1405.0601 (2014)

[109] Zhang, Q., Bhalerao, A., Dickenson, E., Hutchinson, C.: Active appearance pyramids for object parametrisation and fitting. Medical image analysis **32** (2016) 101–114

[110] Zhang, Q., Bhalerao, A., Charles, H.: Deformable appearance pyramids for anatomy representation, landmark detection and pathology classification. International Journal of Computer Assisted Radiology and Surgery (2017)

[111] Moghaddam, B., Pentland, A.: Probabilistic visual learning for object representation. IEEE Transactions on Pattern Analysis and Machine Intelligence **19**(7) (1997) 696–710

[112] Herley, C., Kovacevic, J., Ramchandran, K., Vetterli, M.: Tilings of the time-frequency plane: Construction of arbitrary orthogonal bases and fast tiling algorithms. IEEE Transactions on Signal Processing **41**(12) (1993) 3341–3359

[113] Olivo-Marin, J.C.: Extraction of spots in biological images using multiscale products. Pattern recognition **35**(9) (2002) 1989–1996

[114] Seidenari, L., Serra, G., Bagdanov, A.D., Del Bimbo, A.: Local pyramidal descriptors for image recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on **36**(5) (2014) 1033–1040

[115] Dong, J., Soatto, S.: Domain-size pooling in local descriptors: DSP-SIFT. arXiv preprint arXiv:1412.8556 (2014)

[116] Brox, T., Weickert, J.: A tv flow based local scale measure for texture discrimination. In: European Conference on Computer Vision, Springer (2004) 578–590

[117] Nguyen, M.H., La Torre, F.D.: Local minima free parameterized appearance models. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2008) 1–8

[118] Ashraf, A.B., Lucey, S., Chen, T.: Fast image alignment in the fourier domain. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2010) 2480–2487

[119] Hager, G.D., Belhumeur, P.N.: Efficient region tracking with parametric models of geometry and illumination. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(10) (1998) 1025–1039

[120] Cerrolaza, J.J., Villanueva, A., Cabeza, R.: Shape Constraint Strategies: Novel Approaches and Comparative Robustness. In: Proceedings of the British Machine Vision Conference. (2011) 1–11

[121] Zhang, Q., Bhalerao, A., Helm, E., Hutchinson, C.: Active Shape Model Unleashed with multi-scale local appearance. In: IEEE International Conference on Image Processing. IEEE (2015)

[122] Zhang, S., Zhan, Y., Dewan, M., Huang, J., Metaxas, D.N., Zhou, X.S.: Sparse shape composition: A new framework for shape prior modeling. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2011) 1025–1032

[123] Zhang, S., Zhan, Y., Dewan, M., Huang, J., Metaxas, D.N., Zhou, X.S.: Towards robust and effective shape modeling: Sparse shape composition. Medical Image Analysis **16**(1) (2012) 265–277

[124] Gu, L., Kanade, T.: A generative shape regularization model for robust face alignment. In: Proceedings of the European Conference on Computer Vision. Springer (2008) 413–426

[125] Seiler, C., Pennec, X., Reyes, M.: Capturing the multiscale anatomical shape variability with polyaffine transformation trees. Medical Image Analysis **16**(7) (2012) 1371–1384

[126] Cerrolaza, J.J., Reyes, M., Summers, R.M., González-Ballester, M.Á., Linguraru, M.G.: Automatic multi-resolution shape modeling of multi-organ structures. Medical image analysis (2015)

[127] Popovic, A., de la Fuente, M., Engelhardt, M., Radermacher, K.: Statistical validation metric for accuracy assessment in medical image segmentation. International Journal of Computer Assisted Radiology and Surgery **2**(3-4) (2007) 169–181

[128] Zhang, Q., Bhalerao, A., Parsons, C., Helm, E., Hutchinson, C.: Wavelet appearance pyramids for landmark detection and pathology classification: application to lumbar spinal stenosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2016) 274–282

[129] Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. International Journal of Computer Vision **107**(2) (2014) 177–190

[130] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2) (2004) 91–110

[131] Tzimiropoulos, G., Pantic, M.: Gauss-newton deformable part models for face alignment in-the-wild. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE (2014) 1851–1858

[132] Zhang, Q., Bhalerao, A.: Loglet sift for part description in deformable part models: application to face alignment. Proceedings of BMVC 2016 (2016)

[133] Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: European Conference on Computer Vision, Springer (2012) 679–692

[134] : 300-w faces in-the-wild challenge. `http://ibug.doc.ic.ac.uk/resources/facial-point-annotations/`

[135] Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. IEEE Trans. on Pattern Analysis and Machine Intelligence **35**(12) (2013) 2930–2940

[136] iBUG: Intelligent behaviour understanding group. `https://ibug.doc.ic.ac.uk/`

[137] Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: Database and results. Image and Vision Computing **47** (2016) 3–18

[138] Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2013) 397–403

[139] Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: A semi-automatic methodology for facial landmark annotation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2013) 896–903

[140] Burgos-Artizzu, X.P., Perona, P., Dollár, P.: Robust face landmark estimation under occlusion. In: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE (2013) 1513–1520

[141] Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment at 3000 FPS via regressing local binary features. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE (2014) 1685–1692